BIG DATA CHALLENGES AND ADVANCED COMPUTING SOLUTIONS

JOINT HEARING

BEFORE THE

SUBCOMMITTEE ON ENERGY &
SUBCOMMITTEE ON RESEARCH AND TECHNOLOGY
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
HOUSE OF REPRESENTATIVES

ONE HUNDRED FIFTEENTH CONGRESS

SECOND SESSION

JULY 12, 2018

Serial No. 115–69

Printed for the use of the Committee on Science, Space, and Technology

Available via the World Wide Web: http://science.house.gov

U.S. GOVERNMENT PUBLISHING OFFICE
30–879PDF WASHINGTON : 2018

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

HON. LAMAR S. SMITH, Texas, *Chair*

FRANK D. LUCAS, Oklahoma
DANA ROHRABACHER, California
MO BROOKS, Alabama
RANDY HULTGREN, Illinois
BILL POSEY, Florida
THOMAS MASSIE, Kentucky
RANDY K. WEBER, Texas
STEPHEN KNIGHT, California
BRIAN BABIN, Texas
BARBARA COMSTOCK, Virginia
BARRY LOUDERMILK, Georgia
RALPH LEE ABRAHAM, Louisiana
GARY PALMER, Alabama
DANIEL WEBSTER, Florida
ANDY BIGGS, Arizona
ROGER W. MARSHALL, Kansas
NEAL P. DUNN, Florida
CLAY HIGGINS, Louisiana
RALPH NORMAN, South Carolina
DEBBIE LESKO, Arizona

EDDIE BERNICE JOHNSON, Texas
ZOE LOFGREN, California
DANIEL LIPINSKI, Illinois
SUZANNE BONAMICI, Oregon
AMI BERA, California
ELIZABETH H. ESTY, Connecticut
MARC A. VEASEY, Texas
DONALD S. BEYER, JR., Virginia
JACKY ROSEN, Nevada
CONOR LAMB, Pennsylvania
JERRY McNERNEY, California
ED PERLMUTTER, Colorado
PAUL TONKO, New York
BILL FOSTER, Illinois
MARK TAKANO, California
COLLEEN HANABUSA, Hawaii
CHARLIE CRIST, Florida

SUBCOMMITTEE ON ENERGY

HON. RANDY K. WEBER, Texas, *Chair*

DANA ROHRABACHER, California
FRANK D. LUCAS, Oklahoma
MO BROOKS, Alabama
RANDY HULTGREN, Illinois
THOMAS MASSIE, Kentucky
STEPHEN KNIGHT, California
GARY PALMER, Alabama
DANIEL WEBSTER, Florida
NEAL P. DUNN, Florida
RALPH NORMAN, South Carolina
LAMAR S. SMITH, Texas

MARC A. VEASEY, Texas, *Ranking Member*
ZOE LOFGREN, California
DANIEL LIPINSKI, Illinois
JACKY ROSEN, Nevada
JERRY McNERNEY, California
PAUL TONKO, New York
BILL FOSTER, Illinois
MARK TAKANO, California
EDDIE BERNICE JOHNSON, Texas

SUBCOMMITTEE ON RESEARCH AND TECHNOLOGY

HON. BARBARA COMSTOCK, Virginia, *Chair*

FRANK D. LUCAS, Oklahoma
RANDY HULTGREN, Illinois
STEPHEN KNIGHT, California
BARRY LOUDERMILK, Georgia
DANIEL WEBSTER, Florida
ROGER W. MARSHALL, Kansas
DEBBIE LESKO, Arizona
LAMAR S. SMITH, Texas

DANIEL LIPINSKI, Illinois, *Ranking Member*
ELIZABETH H. ESTY, Connecticut
JACKY ROSEN, Nevada
SUZANNE BONAMICI, Oregon
AMI BERA, California
DONALD S. BEYER, JR., Virginia
EDDIE BERNICE JOHNSON, Texas

CONTENTS

July 12, 2018

Page

Appendix I: Answers to Post-Hearing Questions

Appendix II: Additional Material for the Record

BIG DATA CHALLENGES AND ADVANCED COMPUTING SOLUTIONS

THURSDAY, JULY 12, 2018

HOUSE OF REPRESENTATIVES,
SUBCOMMITTEE ON ENERGY AND
SUBCOMMITTEE ON RESEARCH AND TECHNOLOGY,
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY,
Washington, D.C.

The Subcommittees met, pursuant to call, at 10:15 a.m., in Room 2318, Rayburn House Office Building, Hon. Randy Weber [Chairman of the Subcommittee on Energy] presiding.

LAMAR S. SMITH, Texas
CHAIRMAN

EDDIE BERNICE JOHNSON, Texas
RANKING MEMBER

Congress of the United States
House of Representatives
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
2321 RAYBURN HOUSE OFFICE BUILDING
WASHINGTON, DC 20515–6301
(202) 225–6371
www.science.house.gov

Big Data Challenges and Advanced Computing Solutions

Thursday, July 12, 2018
10:00 a.m.
2318 Rayburn House Office Building

Witnesses

Dr. Bobby Kasthuri, Researcher, Argonne National Laboratory; Assistant Professor, The University of Chicago

Dr. Katherine Yelick, Associate Laboratory Director for Computing Sciences, Lawrence Berkeley National Laboratory; Professor, The University of California, Berkeley

Dr. Matthew Nielsen, Principal Scientist, Industrial Outcomes Optimization, GE Global Research

Dr. Anthony Rollett, U.S. Steel Professor of Materials Science and Engineering, Carnegie Mellon University

U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

HEARING CHARTER

July 12, 2018

TO: Members, Subcommittee on Energy, Subcommittee on Research and Technology

FROM: Majority Staff, Committee on Science, Space, and Technology

SUBJECT: Joint Subcommittee Hearing: "Big Data Challenges and Advanced Computing Solutions"

The Subcommittees on Energy and Research and Technology will hold a hearing titled *Big Data Challenges and Advanced Computing Solutions*, on Thursday, July 12, 2018 at 10:00 a.m. in Room 2318 of the Rayburn House Office Building.

Hearing Purpose:

The purpose of the hearing is to explore the impact of innovative machine learning-based approaches to big data science challenges at the Department of Energy (DOE), in academia, and industry. This hearing will also address the development of these applications within the context of the DOE's mission goals in high performance computing.

Witness List

- **Dr. Bobby Kasthuri**, Researcher, Argonne National Laboratory; Assistant Professor, The University of Chicago
- **Dr. Katherine Yelick**, Associate Laboratory Director for Computing Sciences, Lawrence Berkeley National Laboratory; Professor, The University of California, Berkeley
- **Dr. Matthew Nielsen**, Principal Scientist, Industrial Outcomes Optimization, GE Global Research
- **Dr. Anthony Rollett,** U.S. Steel Professor of Materials Science and Engineering, Carnegie Mellon University

Staff Contact

For questions related to the hearing, please contact Hillary O'Brien of the Majority Staff at 202-226-8984.

Chairman WEBER. The Committee on Science, Space, and Technology will come to order.

Without objection, the Chair is authorized to declare recess of the Subcommittees at any time.

Good morning, and welcome to today's hearing entitled "Big Data Challenges and Advanced Computing Solutions." I now recognize myself for five minutes for an opening statement.

Today, we will explore the application of machine-learning-based algorithms to big-data science challenges. Born from the artificial intelligence—AI—movement that began in the 1950s, machine learning is a data-analysis technique that gives computers the ability to learn directly from data without being explicitly programmed.

Generally speaking—and don't worry; I'll save the detailed description for you all, our expert witnesses—machine learning is used when computers are trained—more than husbands are trained, right, ladies—on large data sets to recognize patterns in that data and learn to make future decisions based on these observations.

Today, specialized algorithms termed "deep learning" are leading the field of machine-learning-based approaches. These algorithms are able to train computers to perform certain tasks at levels that can exceed human ability. Machine learning also has the potential to improve computational science methods for many big-data problems.

As the Nation's largest federal sponsor of basic research in the physical sciences with expertise in big-data science, advanced algorithms, data analytics, and high-performance computing, the Department of Energy is uniquely equipped to fund robust fundamental research in machine learning. The Department also manages the 17 DOE national labs and 27 world-leading scientific user facilities, which are instrumental to connecting basic science and advanced computing.

Machine learning and other advanced computing processes have broad applications in the DOE mission space from high energy physics to fusion energy sciences to nuclear weapons development. Machine learning also has important applications in academia and industry. In industry, common examples of machine-learning techniques are in automated driving, facial recognition, and automated speech recognition.

At Rice University near my home district, researchers seek to utilize machine-learning approaches to address challenges in geological sciences. In addition, the University of Houston's Solutions Lab supports research that will use machine learning to predict the behavior of flooding events and aid in evacuation planning. This would be incredibly beneficial for my district and all areas that are prone to hurricanes and to flooding. In fact, in Texas we're still recovering from Hurricane Harvey, the wettest storm in United States history.

The future of scientific discovery includes the incorporation of advanced data analysis techniques like machine learning. With the next generation of supercomputers, including the exascale computing systems that DOE is expected to field by 2021, American researchers utilizing these technologies will be able to explore even bigger challenges. With the immense potential for machine-learning technologies to answer fundamental scientific questions, provide the foundation for high-performance computing capabilities, and to drive future technological development, it's clear that we should prioritize this research.

I want to thank our accomplished panel of witnesses for their testimony today, and I look forward to hearing what role Congress should play in advancing this critical area of research.

[The prepared statement of Chairman Weber follows:]

COMMITTEE ON
SCIENCE, SPACE, & TECHNOLOGY
Lamar Smith, Chairman

For Immediate Release
July 12, 2018

Media Contacts: Heather Vaughan, Bridget Dunn
(202) 225-6371

Statement by Chairman Randy Weber (R-Texas)

Big Data Challenges and Advanced Computing Solutions

Chairman Weber: Good morning and welcome to today's joint Energy and Research and Technology Subcommittee hearing. Today, we will explore the application of machine learning–based algorithms to big data science challenges.

Born from the Artificial Intelligence (AI) movement that began in the 1950s, machine learning is a data analysis technique that gives computers the ability to learn directly from data without being explicitly programmed.

Generally speaking, and don't worry I'll save the detailed description for our expert witnesses, machine learning is used when computers are "trained" on large data sets to recognize patterns in that data, and learn to make future decisions based on these observations.

Today, specialized algorithms termed "deep learning" are leading the field of machine learning-based approaches. These algorithms are able to train computers to perform certain tasks at levels that can exceed human ability. Machine learning also has the potential to improve computational science methods for many big data problems.

As the nation's largest federal sponsor of basic research in the physical sciences, with expertise in big data science, advanced algorithms, data analytics and high performance computing, the Department of Energy (DOE) is uniquely equipped to fund robust fundamental research in machine learning.

The Department also manages the 17 DOE national laboratories and 27 world-leading scientific user facilities, which are instrumental to connecting basic science and advanced computing.

Machine learning and other advanced computing processes have broad applications in the DOE mission space: from high energy physics to fusion energy sciences to nuclear weapons development.

Machine learning also has important applications in academia and industry. In industry, common examples of machine learning techniques are in automated driving, facial recognition and automated speech recognition.

At Rice University near my home district, researchers seek to utilize machine learning approaches to address challenges in geological sciences. In addition, the University's

Houston Solutions Lab supports research that will use machine learning to predict the behavior of flooding events and aid in evacuation planning. This would be incredibly beneficial for my district and all areas prone to hurricanes and flooding. In Texas, we are still recovering from Hurricane Harvey—the wettest storm on record!

The future of scientific discovery includes the incorporation of advanced data analysis techniques like machine learning.

With the next generation of supercomputers, including the exascale computing systems that DOE is expected to field by 2021, American researchers utilizing these technologies will be able to explore even bigger challenges.

With the immense potential for machine learning technologies to answer fundamental scientific questions, provide the foundation for high performance computing capabilities and drive future technological development, it's clear we should prioritize this research.

I want to thank our accomplished panel of witnesses for their testimony today and I look forward to hearing what role Congress should play in advancing this critical area of research.

###

Chairman WEBER. I now recognize the Ranking Member for an opening statement.

Mr. VEASEY. Thank you, Chairman Weber. Thank you, Chairwoman Comstock, and also, thank you to the distinguished panel for being here this morning.

As you know, there are a growing number of industries today that are relying on generating and interpreting large amounts of data to overcome new challenges. The new—the energy sector in particular is making strides in leveraging these new technologies and techniques. Today, we're going to hear more about the advancements that we're going to see in the upcoming years.

Sensor-equipped aircraft engines, locomotive, gas, and wind turbines are now able to track production efficiency and the wear and tear on vital machinery. This enables significant reductions in fuel consumption, as well as carbon emissions. The technologies are also significantly improving our ability to detect failures before they occur and prevent disasters, and by doing so will save money, will save time, and lives. And by using analytics, sensors, and operational data, we can manage and optimize systems ranging from energy storage components to power plants and to the electric grid.

As digital technologies revolutionize the energy sector, we also must ensure the safe and responsible use of these processes. With our electric grid always in under persistent threats from everything from cyber to other modes of subterfuge, the security of these connected systems is of the utmost importance. Nevertheless, I'm excited to learn more about the value and benefits that these technologies may be able to provide for our economy and our environment alike.

I'm looking forward to hearing what we can do in Congress to help guide and support the responsible development of these new data-driven approaches to the management of these evermore complex systems that our society is very dependent on.

Thank you, and, Mr. Chairman, I yield back the balance of my time.

[The prepared statement of Mr. Veasey follows:]

<u>OPENING STATEMENT</u>
Ranking Member Marc Veasey (D-TX)
of the Subcommittee on Energy

House Committee on Science, Space, and Technology
Subcommittee on Energy
Subcommittee on Research and Technology
"Big Data Challenges and Advanced Computing Solutions"
July 12, 2018

Thank you, Chairman Weber and Chairwoman Comstock for holding this hearing today, and thank you to this excellent panel of witnesses for being here this morning.

A growing number of industries today are relying on generating and interpreting large amounts of data to overcome new challenges. The energy sector in particular is making strides in leveraging these new technologies and techniques.

Today, we'll hear more about the advancements we'll see in the coming years. Sensor-equipped aircraft engines, locomotives, gas turbines, and wind turbines are now able to track production efficiency and the wear and tear on vital machinery. This enables significant reductions in fuel consumption as well as carbon emissions.

The technologies are also significantly improving our ability to detect failures before they occur and prevent disasters. By doing so, we save money, time, and lives. By using analytics, sensors, and operational data, we can manage and optimize systems ranging from energy storage components to power plants to the electric grid.

As digital technologies revolutionize the energy sector, we also must ensure the safe and responsible use of these processes. With our electric grid under persistent cyber threats, the security of these connected systems is of the utmost importance. Nevertheless, I am excited to learn more about valuable benefits that these technologies may be able to provide for our economy and our environment alike.

I look forward to learning about what we in Congress can do to guide and support the responsible development of these new data-driven approaches to the management of the ever-more-complex systems that our society now depends on.

Thank you, and I yield back the remainder of my time.

Chairman WEBER. Thank you, Mr. Veasey.

I now recognize the Chairwoman of the Research and Technology Subcommittee, the gentlewoman from Virginia, Mrs. Comstock, for an opening statement.

Mrs. COMSTOCK. Thank you, Chairman Weber.

A couple of weeks ago, our two Subcommittees joined together on a hearing to examine the state of artificial intelligence and the types of research being conducted to advance this technology. The Committee learned about the nuances of the term artificial intelligence, such as the difference between narrow and general AI and implications for a world in which AI is ubiquitous.

Today, we delve deeper into disciplines originating from the AI movement of the 1950s that include machine learning, deep learning, and neural networks. Until recently, machine learning and especially deep-learning technologies were only theoretical because deep-learning models require massive amounts of data and computing power. But advances in high-performance graphics, processing units, cloud computing, and data storage have made these techniques possible.

Machine learning is pervasive in our day-to-day lives from tagging photos on Facebook to protecting emails with spam filters to using a virtual assistant like Siri or Alexa for information. Machine-learning-based algorithms have powerful applications that ultimately help make our lives more fun, safe, and informative.

In the federal government, the Department of Energy stands out for its work in high-performance computing and approaches to big-data science challenges. The Energy Department researchers are using machine-learning approaches to study protein behavior, to understand the trajectories of patient health outcomes, and to predict biological drug responses. At Argonne National Laboratory, for example, researchers are using intensive machine-learning-based algorithms to attempt to map the human brain.

A program of particular interest to me involves a DOE and Department of Veterans Affairs venture known as the MVP–CHAMPION program. This joint collaboration will leverage DOE's high-performance computing and machine-learning capabilities to analyze health records of more than 20 million veterans maintained by the VA. The goal of this partnership is to arm the VA with data it can use to potentially improve health care offered to our veterans by developing new treatments and preventive strategies and best practices.

The potential for AI to help humans and further scientific discoveries is obviously immense. I look forward to what our witnesses will testify to today about their work and—which may give us a glimpse into the revolutionary technologies of tomorrow that we're here to discuss.

So I thank you, Mr. Chairman, and I yield back.

[The prepared statement of Mrs. Comstock follows:]

COMMITTEE ON
SCIENCE, SPACE, & TECHNOLOGY
Lamar Smith, Chairman

For Immediate Release
July 12, 2018

Media Contacts: Heather Vaughan, Bridget Dunn
(202) 225-6371

Statement by Chairwoman Barbara Comstock (R-Va.)

Big Data Challenges and Advanced Computing Solutions

Chairwoman Comstock: A couple of weeks ago, our two subcommittees joined together on a hearing to examine the state of artificial intelligence (AI) and the types of research being conducted to advance this technology. The committee learned about the nuances of the term artificial intelligence—such as the difference between narrow and general AI—and implications for a world in which AI is ubiquitous. Today, we delve deeper into disciplines originating from the AI movement of the 1950s that include machine learning, deep learning and neural networks.

Until recently, machine learning and especially deep learning techniques were only theoretical, because deep learning models require massive amounts of data and computing power. But advances in high performance graphics processing units, cloud computing and data storage have made these techniques possible.

Machine learning is pervasive in our day to day lives—from tagging photos on Facebook, to protecting emails with spam filters, to using a virtual assistant like Siri or Alexa for information—machine learning-based algorithms have powerful applications that ultimately help make our lives more fun, safe and informative.

In the federal government, the Department of Energy (DOE) stands out for its work in high performance computing and approaches to big data science challenges. DOE researchers are using machine learning approaches to study protein behavior, to understand the trajectories of patient health outcomes and to predict biological drug responses. At Argonne National Laboratory for example, researchers are using intensive machine learning-based algorithms to attempt to map the human brain!

A program of particular interest to me involves a DOE and Department of Veterans Affairs (VA) venture known as the MVP-CHAMPION program.

This joint collaboration will leverage DOE's high performance computing and machine learning capabilities to analyze health records of more than 20 million veterans maintained by the VA. The goal of this partnership is to arm the VA with data it can use to potentially improve health care offered to veterans by developing new treatments and preventive strategies.

The potential for AI to help humans and further scientific discoveries is immense. I look forward to what our witnesses have to say about their work today—which may give us a glimpse into the revolutionary technologies of tomorrow.

###

Chairman WEBER. I thank the gentlelady.

And let me introduce our witnesses. Our first witness is Dr. Bobby—Mr. Chairman, are you going to——

Chairman SMITH. Mr. Chairman, thank you. In the interest of time, I just ask unanimous consent to put my opening statement in the record.

Chairman WEBER. Without objection.

[The prepared statement of Chairman Smith follows:]

COMMITTEE ON
SCIENCE, SPACE, & TECHNOLOGY
Lamar Smith, Chairman

For Immediate Release
July 12, 2018

Media Contacts: Heather Vaughan, Bridget Dunn
(202) 225-6371

Statement by Chairman Lamar Smith (R-Texas)

Big Data Challenges and Advanced Computing Solutions

Chairman Smith: Today we will hear from a panel of experts on a number of big data challenges facing the Department of Energy (DOE), academia, and industry, and the innovative computing approaches used to address them.

Recent advances in our ability to store and process information have led to a growth of large and complex data sets. At the same time, greater computing power and increasingly sophisticated algorithms have allowed for dramatic advances in artificial intelligence and machine learning. These tools have powerful applications for challenges with a large amount of data on which to train computing systems.

Machine learning is a practice in which computers not only analyze data, but then use that analysis and data to refine and enhance future predictions. Essentially, it gives computers the ability to learn directly from data without being explicitly programmed.

This advanced technology is already creating tremendous developments in many fields including medicine, manufacturing and finance.

Whether it's protecting your credit card from fraudulent activity to helping you find the fastest way to work, we all benefit from machine learning every day.

Machine learning is especially valuable when analyzing big data. As the nation's largest federal supporter of basic research in the physical sciences, DOE is well suited to develop and apply machine learning across its research portfolio.

DOE funds robust programs in advanced scientific computing and applied mathematics, and hosts the fastest supercomputers in the world at DOE national labs. The Department also funds research in a wide range of scientific disciplines—from physics and chemistry, to materials science and biology.

DOE has a specific research need to address big data challenges and is uniquely positioned to advance machine learning-based approaches to solving these challenges.

For example, machine learning-based algorithms have the ability to revolutionize material science research. The discovery of new materials has been instrumental to many recent advancements in carbon capture, battery and solar cell technologies. At Lawrence Berkeley National Laboratory and at SLAC National Accelerator Laboratory, researchers are utilizing machine learning-based approaches to shorten the timeline of the materials discovery process.

Machine learning is also particularly useful in the biological and biomedical sciences. In many of these areas, like the study of microbial data, the behavior of proteins, and even patient care, we have the potential to make significant scientific progress by using detailed analysis of large amounts of data.

At Argonne National Laboratory, researchers have a plan to create a 3D map of neurons in the human brain. By utilizing the imaging power of the Advanced Photon Source, and the leadership computing facility at Argonne, researchers can collect and fit together millions of high resolution images of mammal brains to reconstruct their complex structures and characterize their behavior. I look forward to hearing more about this exciting area of research today.

American universities are also taking advantage of machine learning-based approaches to big data challenges. At Carnegie Mellon University's NextManufacturing Center, researchers have focused on how to combine 3D printing and machine learning to monitor the quality of manufactured components in real-time.

These are just a few of the issues already being addressed by machine learning. Continued development will allow us to address more complex challenges and advance scientific discovery.

With new exascale and quantum computing systems, more big data challenges will be within our reach. We must continue to support the research in applied mathematics and computer science that will help develop the next generation of computing tools.

I thank the witnesses for their testimony and look forward to a valuable discussion of this important science today.

###

[The prepared statement of Ranking Member Johnson follows:]

OPENING STATEMENT
Ranking Member Eddie Bernice Johnson (D-TX)

House Committee on Science, Space, and Technology
Subcommittee on Energy
Subcommittee on Research and Technology
"Big Data Challenges and Advanced Computing Solutions"
July 12, 2018

I'd like to thank Chairman Weber, Ranking Member Veasey, Chairwoman Comstock, and Ranking Member Lipinski for holding this hearing, and thank you to our witnesses for being here this morning.

As highlighted in our previous Committee hearing on this topic last month, artificial intelligence has potentially powerful applications for a wide range of industries. In the energy sector, these technologies are currently gaining traction by providing efficient and innovative ways to optimize the use, production, and distribution of energy resources.

When many people think of artificial intelligence, they think of science fiction movies and technologies that are far in the future, but in truth, they are already playing a significant role in our lives today. With the rise of what is being called big data, artificial intelligence is playing an even more important role. The amount of data available is quickly becoming far too large to be handled by human workforces in a timely fashion, creating the need for machine driven solutions.

One of the topics we are discussing today is called Machine Learning. Machines are being trained to be able to "learn" from data, and then automatically perform meaningful tasks based on that analysis. There are many potential benefits to machine learning. It will simplify and expedite many processes as well as improve safety across various sectors. And it will allow our workforce to focus on more critical thought-based problems that a computer simply can't do. While STEM education is not a focus of this hearing, the hearing topic does remind us of the critical importance of improving STEM learning and access to quality STEM education at all levels. The kinds of good-paying jobs that were once a ticket into the middle class with just a high school diploma are going away, and artificial intelligence is one factor in this changing economy.

Attention to the benefits and risks of artificial intelligence can be seen across private industry, government agencies, and in academic research. With the ability to enhance and streamline energy production, utility companies look to these technologies as a way to increase efficiency as well as safety throughout their operations. As we'll hear more about in the testimony we receive this morning, agencies such as the Department of Energy are incorporating machine learning into materials and biomedical research through our universities and national laboratories. Its use offers the promise that it can lead to previously unattainable innovations that will save us time, money, and even lives in the not-too-distant future.

However, as more data comes in, we must also ask where this data is coming from and what the risks may be as we increasingly rely on machine learning. For example, we must consider how the data is distributed, processed, analyzed. And – when the data includes sensitive information relevant to our security or the privacy of our citizens – how it is being protected. Other questions policy makers must ask are who is benefitting from this data, and what concerns should we have about the amount and quality of data being produced. Finally, as I mentioned previously, we must consider the workforce that will be necessary to take advantage of, as well as mitigate the risk of all aspects of big data. These are questions that I'm sure our witnesses today can provide some further insight into. It is our duty within this Committee to not only examine how these technologies will benefit us, but to also contemplate what new challenges will emerge as well.

I look forward to learning more about what we in Congress can do to responsibly support the development and use of these breakthrough technologies. Lastly, before I close, I would like to welcome the visiting interns from the American Institute of Physics who are here with us this morning. I hope you all enjoy this experience, and it inspires you to stay engaged throughout your careers on the many important science policy issues that we'll all need your help in addressing. And with that I yield back.

[The prepared statement of Mr. Lipinski follows:]

OPENING STATEMENT

Ranking Member Daniel W. Lipinski (D-IL)
of the Subcommittee on Research and Technology

House Committee on Science, Space, and Technology
Subcommittee on Energy
Subcommittee on Research and Technology
"Big Data Challenges and Advanced Computing Solutions"
July 12, 2018

Thank you, Chairman Weber and Chairwoman Comstock, for holding this hearing to explore the impact of machine learning-based approaches to big data science challenges at the Department of Energy, in academia, and in industry.

During a hearing last month, this committee heard from expert witnesses about the state of artificial intelligence and machine learning technology. That hearing was an opportunity to understand the history of AI and machine learning and their current and future impact on society, including jobs, the economy, and workforce needs. Today's witnesses will expand on machine learning solutions for challenges faced by the energy industry.

The energy industry is turning to applications of machine learning to help improve power generation, transmission and distribution, exploration of oil and gas resources, and materials characterization. Companies such as GE are already using data-driven predictive analytics to reduce their fuel consumption and lower their carbon footprint. The data produced by sensors and analyzed by sophisticated software allow for better matching of supply and demand, more efficient operation of the grid, and better integration of new technologies such as renewable energy generation and electric vehicles. In addition to the private sector, the federal government has made longstanding investments in artificial intelligence and data science research to grow our national machine learning capabilities, many of which can be applied to energy grid resiliency efforts. Experts have warned of the disastrous consequences of a natural or man-made attack on the grid. Through its Grid Modernization Initiative, the Department of Energy is working with public and private sector partners to develop technologies, including big data and machine learning, needed to meet current and future demands on the energy grid. And through the National Labs, represented today by Dr. Kasthuri from Argonne National Lab in my district, the Department of Energy has developed some of the world's foremost high-performance computing infrastructure to support advancing the frontiers of data science.

As the energy industry increases its use of big data and machine learning, we must consider the appropriate balance and scale of federal support. Advanced computing solutions increase the usability of the large amounts of data produced by the energy sector which can help achieve more efficient production, providing broad societal benefit. However, there are still technical areas to be addressed including labeling and sharing of data, bias, confidence in output, and other issues.

I stated during last month's hearing that the Science Committee and our other colleagues here in Congress have a responsibility to inquire about the technical issues as well as the societal and economic impacts of AI and machine learning. Ensuring a skilled workforce for the machine learning and AI-based jobs of tomorrow is a high priority because there is a global race to assert leadership in AI. The U.S. must leverage its role as an incubator of ingenuity and innovation to be at the forefront of this technology.

I thank all of the witnesses for being here today and look forward to learning how Congress can help improve the use of machine learning and AI technologies to address big data science challenges in the energy sector.

I yield back.

Chairman WEBER. Thank you. I appreciate that.

Now, I will introduce the witnesses. Our first witness is Dr. Bobby Kasthuri, the first neuroscience researcher at Argonne National Lab and an Assistant Professor in the Department of Neurobiology at the University of Chicago. You're busy. Dr. Kasthuri's current research focuses on innovation and new approaches to brain mapping, including the use of high-energy x-rays from synchrotron sources for mapping brains in their entirety.

He holds a Bachelor of Science from Princeton University, an M.D. from Washington University School of Medicine, and a Ph.D. from Oxford University where he studied as a Rhodes scholar. Welcome, Doctor.

Our second witness today is Dr. Katherine Yelick, a Professor of Electrical Engineering and Computer Sciences at the University of California, Berkeley, and the Associate Laboratory Director for Computing at Lawrence Berkeley National Laboratory. Her research is in high-performance computing, programming languages, compilers, parallel algorithms, and automatic performance tuning.

Dr. Yelick received her Bachelor of Science, Master of Science, and Ph.D. all in computer science at the Massachusetts Institute of Technology. Welcome, Dr. Yelick.

Our next witness is Dr. Matthew Nielsen, Principal Scientist at the GE Global Research Center. Dr. Nielsen's current research focuses on digital twin and computer modeling and simulation of physical assets using first-principle physics and machine-learning methods.

He received a Bachelor of Science in physics at Alma College in Alma, Michigan, and a Ph.D. in applied physics from Rensselaer.

Dr. NIELSEN. Rensselaer.

Chairman WEBER. Rensselaer, okay, Polytechnic Institute in Troy, New York. Welcome, Dr. Nielsen.

And our final witness today is Dr. Anthony Rollett, the U.S. Steel Professor of Metallurgical Engineering and Materials Science at Carnegie Mellon University, a.k.a. CMU. Dr. Rollett has been a Professor of Materials Science Engineering at CMU for over 20 years and is the Co-Director of CMU's NextManufacturing Center. Dr. Rollett's research focuses on microstructural evolution and microstructure property relationships in 3–D.

He received a Master of Arts in metallurgy and materials science from Cambridge University and a Ph.D. in materials engineering from Drexel University. Welcome, Dr. Rollett.

I now recognize Dr. Kasthuri for five minutes to present his testimony. Doctor?

TESTIMONY OF DR. BOBBY KASTHURI, RESEARCHER, ARGONNE NATIONAL LABORATORY; ASSISTANT PROFESSOR, THE UNIVERSITY OF CHICAGO

Dr. KASTHURI. Thank you. Chairman Smith, Chairman Weber, Chairwoman Comstock, Ranking Members Veasey and Lipinski, and Members of the Subcommittees, thank you for this opportunity to talk and appear before you. My name is Bobby Kasthuri. I'm a Neuroscientist at Argonne National Labs and an Assistant Professor in the Department of Neurobiology at the University of Chicago.

And the reason I'm here talking to you today is because I think we are at a pivotal moment in our decades-long quest to understand the brain. And the reason we're at this pivotal moment is that we're actually witnessing in real time is the collision of two different disciplines, two different worlds, the worlds of computer science and neuroscience. And if we can nurture and develop this union, it could fundamentally change many things about our society.

First, it could fundamentally change how we think about understanding the brain. It could change and revolutionize how we treat mental illness, and perhaps even more significantly, it can change how we think and imagine and build our future computers and our future robots based on how brains solve problems.

The major obstacle between us and realizing this vision is that, for many neuroscientists, modern neuroscience is extremely expensive and extremely resource-intensive. To give you an idea of the scale, I thought it might help to give you an example of the enormity of the problem that we're trying to do.

The human brain, your brains, probably contain on order 100 billion brain cells or neurons, and the main thing that neurons do is connect with each other. And so in your brain there's probably—each neuron connects on average 10,000 times with 10,000 other neurons. That means in your brain there are orders of magnitude more connections between neurons than stars in the Milky Way galaxy. And what's even more important for neuroscientists is that we believe that this map, this map of you, this map of connections contains all of the things that make us human. Our creativity, our ability to think critically, our fears, our dreams are all contained in that map.

But unfortunately, that map, if we were to do it, wouldn't be one gigabyte of data; it wouldn't be 100 gigabytes of data. It could be on order a billion gigabytes of data, perhaps the largest data set about anything ever collected in the history of humanity. The problem is that for many neuroscientists even analyzing a fraction of this map is beyond their resources, the resources of their laboratory, the resources of the universities, and perhaps the resources of even large institutions. And if we don't address this gap, then what will happen is that only the richest neuroscientists will be able to answer their questions, and we would like every neuroscientist to have access to answer the most important questions about brains and ultimately promote this fusion of computer science and neuroscience.

Luckily, there is a potential solution, and the potential solution is the Department of Energy and the national lab system, which is part of the Department of Energy. As stewards of our scientific architecture, as stewards of some of the most advanced technological and computing capabilities available, the Department of Energy and the national labs can address this gap, and in fact, they do address this gap in many different sciences.

If I was a young astrophysicist or a young materials scientist, no one would expect me to get money and build my own space telescope. Instead, I would leverage the amazing resources of the national lab system to answer my fundamental questions. And although many fields of science have learned how to leverage the expertise and the resources available in the national lab system, neuroscientists have not.

A national center for brain mapping situated within the DOE lab system could actually be a sophisticated clearinghouse to ensure that the correct physics and engineering and computer science tools are vetted and accessible for measuring brain structure and brain function. Since the national labs are also the stewards of our advanced computing infrastructure, they're ideally suited to incubate these revolutions in computer and neurosciences.

Decades earlier, as a biologist, I just recently learned that the DOE and the national labs helped usher in humanity's perhaps greatest scientific achievement of the 20th century, the mapping of the human genome and the understanding of the genetic basis of life. We believe that the DOE and the national lab system can make a similar contribution to understanding the human brain.

Other countries like Japan, South Korea, and China, cognizant of the remarkable benefits to economic and national security that understanding brains and using them to make computer science better have already invested in national efforts in artificial intelligence and national efforts to understand the brain. The United States has not yet, and I think it's important at the end of my statement for everyone to remember that we are the ones who went to the moon, we are the ones who harnessed the power of nuclear energy, and we are the ones that led the genomic revolution. And I suspect it's the moment now for the United States to lead again, to map and help reverse engineer the physical substrates of human thought, arguably the most challenging quest of the 21st century and perhaps the last great scientific frontier.

Thank you for your time and attention today. I welcome any questions you might have.

[The prepared statement of Dr. Kasthuri follows:]

Written Testimony of Dr. Narayanan (Bobby) Kasthuri
Neuroscientist, Argonne National Laboratory, and
Assistant Professor of Neurobiology, University of Chicago
before the
Committee on Science, Space, and Technology, Subcommittee on Energy and Subcommittee on Research and Technology, of the U.S. House of Representatives
July 12th, 2018

SUMMARY

- We stand at a pivotal moment in our centuries-long quest to understand the brain—the moment when the worlds of computer science and neuroscience collide.
  - We can transform how we treat mental illness and brain diseases.
  - We can revolutionize how we think about and build future computers and algorithms.
  - We can bolster our artificial intelligence capabilities and national and economic security.
- Modern neuroscience is expensive and resource intensive.
  - Researchers encounter both financial and structural barriers to entry; needed investments in physics, engineering and computer science are typically beyond the scope of laboratories at single universities and institutes.
  - With the neuroscience community unable to efficiently utilize current capabilities, we are limiting the types of hypotheses we test to drive the next generation of innovation.
  - We must counteract the widening gap between the small fraction of laboratories utilizing the most recent technology and the remaining majority of neuroscientists.
- The DOE and the national lab system are perfectly suited to address this gap.
  - The national laboratories act as stewards of large-scale infrastructure supporting many of the nation's scientific programs; however, until recently there has been limited interaction between the labs and the neuroscience community.
  - A national clearinghouse will ensure that the necessary physics, engineering and computer science resources are vetted and freely accessible to measure brain structure and functions.
  - As stewards of the nation's advanced computing infrastructure, the labs can support efforts to understand the brain just as they supported mapping the human genome.
- With 100 billion brain-cells (neurons) making an average of 10,000 connections with each other, the human brain is the most complicated structure studied in the history of humanity.
  - Understanding how it functions will be the great intellectual achievement of the 21st century, revealing the physical bases of our most human abilities like reasoning and serving as the blueprint for reverse engineering those abilities into algorithms and robots.
  - Other countries like Japan, South Korea, and China, cognizant of the enormous economic and national security benefits of understanding the brain, have committed national efforts to both brain mapping and artificial intelligence; the United States has not.
- We went to the moon, we harnessed the power of nuclear energy, and we led the genomic revolution—now is the moment for the United States to lead again.
  - By mapping and reverse engineering the physical substrates of human thought, we will complete the most challenging quest of the 21st Century and cross what could be the last great scientific frontier.

Written Testimony of Dr. Narayanan (Bobby) Kasthuri

Neuroscientist, Argonne National Laboratory, and

Assistant Professor of Neurobiology, University of Chicago

before the

Committee on Science, Space, and Technology, Subcommittee on Energy and Subcommittee on Research and Technology, of the U.S. House of Representatives

July 12, 2018

Chairman Weber, Chairwoman Comstock, Ranking Members Veasey and Lipinski, and members of the subcommittees, thank you for this opportunity to appear before you. My name is Bobby Kasthuri, and I am a neuroscience researcher at the U.S. Department of Energy's (DOE's) Argonne National Laboratory and an assistant professor of neurobiology at the University of Chicago.

I am here today because I believe that understanding the human brain is the most challenging quest of the 21st century—perhaps the last great scientific frontier—and that advanced capabilities and facilities of the DOE National Laboratories are critical to help usher in a new era of understanding. Scientists began and ended the great scientific challenge of the previous century—understanding the genetic basis of life—by creating two maps: one of the atomic structure of DNA in 1953, and another of every nucleotide in a human genome in 2003. The science enabled by Watson and Crick and the Human Genome Project is revolutionizing our understanding of the genetic bases of human health and disease.

A similar revolution awaits us when we understand how human brains acquire knowledge from experience—how we find patterns in our senses and use them to plan and act. When we know exactly how those processes work, we can connect prosthetic bodies to the paralyzed, design rational medical treatments for brain disease, and reverse-engineer human cognition into our computers, potentially at the energy cost of a fraction of a common lightbulb.

The medical ramifications alone are tremendous:

- The National Alliance on Mental Illness (NAMI) indicates that approximately 1 in 5 adults in the United States—43.8 million people—experiences mental illness in a given year, resulting in nearly $200 billion in lost earnings annually.

- The Alzheimer's Association reports that an estimated 5.7 million Americans of all ages are living with the disease; it is currently the sixth leading cause of death in the United States. In 2018, Alzheimer's and other dementias will cost the nation $277 billion, with costs rising as high as $1.1 trillion by 2050.

- The Centers for Disease Control and Prevention report that 1 in 59 children have autism-spectrum disorders that cause mild to severe social challenges and communication difficulties, as well as physical and medical issues.

Given the enormous benefits a better understanding of brains could provide, you could ask why we have not made more progress. Part of the problem is the sheer complexity of the human brain. The human brain contains around 100 billion cells, or neurons, which make thousands of connections called synapses with each other. The complexity of this intricate communication web cannot be overstated. Parts of our nervous system beyond our brain, some just mere atoms long, extend from foot to spine. The quest to understand how the brain works requires more cooperation across academic disciplines than any other human endeavor.

The good news is that we have defined the underlying hardware, so to speak. Every nervous system is based on the same principle—all representations, computations and actions mediated by the brain depend on neurons that are connected by synapses in highly complicated directional networks. Each neuron receives information from synapses that connect to its dendrites (branches) and sends information via its axon, which connects to dendrites of other neurons. One neuron might receive thousands of separate messages and convey the integrated information to thousands of other neurons.

You can picture each neuron as a hub that sends and receives signals to and from many thousands of other neurons. Neuroscientists propose that the map of how those 100 billion neurons make 1 quadrillion connections with each other, what we call the "connectome," is a map of who you are: your skills, your memories, your fears, and your personality. Disruptions or alterations in these maps—"mis-wirings" between neurons—are the basis of many neurological and psychiatric disorders.

The "Mind-Meld" Between Computer Science and Neuroscience

In our quest to understand the brain, one of the most important scientific collaborations is the "mind meld" between computer science and neuroscience. Given the complexity of the brain I just described, you can imagine that no matter how neuroscientists analyze the brain—whether we use laser beams, genetic engineering, fluorescent proteins, pharmaceuticals, virtual reality, metamaterials or robotics—tremendous computing power will always be a necessity.

Neuroscience, perhaps more than any other field of biology, operates at the cutting edge of big data. The raw data for the connectome, or map, I described will measure approximately 1 trillion gigabytes (an exabyte) and could not fit in the memory of any current computer. For comparison, the entire Human Genome Project measures only a few gigabytes. Indeed, if you could combine all the written material in the world into one dataset, it would be just a small fraction of the size of this brain map.

Scientists, including those at Argonne National Laboratory and the University of Chicago, and collaborators around the United States, are already working toward a human connectome by mapping smaller brains of other animals. To create even the smallest neural map teams of neuroscientists and computer scientists must work side by side to analyze the enormous brain datasets and use the latest artificial intelligence technology. Interestingly, we have discovered that although this collaboration clearly furthers neuroscience, this work is mutually beneficial to advancing computer science as well.

First, it turns out that problems to which computer scientists are eager to apply artificial intelligence—understanding pedestrian behavior to ensure the safe operation of a self-driving car or automatically interpreting changes in satellite images over time for strategic intelligence—involve the rapid analysis of large datasets at the same scales sought by neuroscientists. The only difference is that brain datasets are already orders of magnitude larger than any datasets humans have ever collected and are guaranteed to grow even larger. Deciphering the human brain by creating a new generation of artificial intelligence that is capable of analyzing the largest datasets ever created will inevitably aid every other field of human endeavor that struggles with big data.

Second, and perhaps even more importantly, understanding the brain more deeply could lead to a revolution in computing. Even as they herald recent gains in the computational abilities of artificial neural networks, computer scientists remain concerned that conventional approaches will soon plateau in performance. Almost every human brain possesses fundamental skills that even the most sophisticated algorithms do not: reasoning, humor, learning and creativity. If we can find the physical bases of these abilities in the brain, we can transform the landscape of computing.

The Future Is Here

Neuroscientists around the world—including a coalition comprised of both researchers from Argonne National Laboratory and collaborators from Princeton University (NJ), Baylor University (TX), Rice University (TX), the University of Notre Dame (IN), the Allen Brain Science Institute (WA), and other U.S. institutions—already have begun trying to reverse-engineer how brains work, to discover uniquely biological algorithms. For example, as part of the IARPA MICrONS program, a component of the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) initiative, neuroscientists seek to reveal fundamental aspects of the brain's learning machinery from simpler animals. By observing the dynamics of the living mouse brain as it learns, and mapping the connections between neurons that mediate the learning, we hope to decipher how the brain uses its hardware in combination with programming language to recognize objects. Scientists will then be able to incorporate those principles into the next generation of computer programs. Artificial intelligence has progressed rapidly, but studying the best computer we know—the brain—has the potential to generate novel networks with leaps in performance that would otherwise take many years of chiseling and searching to achieve.

Indeed, computer scientists used the crudest visual maps of primate brains to develop what would become the ancestors of machine learning and other successful modern artificial intelligence. Given this past success, we expect that increasingly detailed maps of mouse brains will bear the next generation of computer algorithms. At only the halfway point of the 5-year MICrONS project, early results already suggest this historic data will yield countless insights for many years to come. Teams at Princeton, Baylor, Rice, and the Allen Brain Institute already have leveraged cutting-edge machine vision and artificial intelligence algorithms to produce exquisite maps of mouse brains with unprecedented detail, which already are changing the foundations of neuroscience and computer science. However, even the smallest part of the mouse brain, a small fraction in size and capability relative to the human brain, is the limit of most scientists, universities, and institutes. To map the human brain will require scholars with incredibly diverse expertise and skillsets, collaborating with federal scientific agencies like the National Institutes of Health, National Science Foundation, and the DOE and its National Laboratories. It is an interdisciplinary project of great scope and tremendous potential.

A National Resource for Neuroscience and Artificial Intelligence

Although neuroscientists and computer scientists are making remarkable progress, an unfortunate reality still prevents us from fully understanding the human brain and leveraging these discoveries for society—that is, most neuroscientists lack access to the tools and resources needed to test their ideas about the brain. Indeed, the enduring success of the BRAIN initiative will depend on widespread access to the technological advancements, computational tools and datasets the initiative creates.

Today the neuroscience community is underutilizing current technological capabilities, limiting the types of questions and hypotheses we can test to drive the next generation of innovation. We must counteract the widening gap between the small fraction of laboratories utilizing the most recent technology and the remaining majority of neuroscientists. A sophisticated national clearinghouse will ensure that the physics, engineering and computer science are vetted and freely accessible to measure brain structure and functions.

The DOE and the National Laboratory system are uniquely suited to convene leading researchers across the various scientific disciplines to overcome these barriers. At the forefront of discovery and innovation across fundamental sciences, the DOE National Laboratories are stewards of large-scale scientific user facilities, including light sources, accelerators, and supercomputing facilities that support advancements in a range of disciplines from astrophysics to chemistry to material science; however, until recently interaction between the neuroscience community and the National Laboratory system has been limited. Indeed, a Secretary of Energy Advisory Board (SEAB) reported to the DOE this exact sentiment (Secretary of Energy Advisory Board Report of the Task Force on Biomedical Sciences, September 22, 2016, p. 14)

"Brain research is supported across many institutes of the NIH, but the opportunities for DOE involvement are perhaps best appreciated in the context of the recent BRAIN Initiative. ... BRAIN has begun a concerted effort to improve the methods available for brain research, both for experimental work and in the domain of theory and analysis. The ultimate goal is to understand large circuits of nerve cells: What are all the types of neurons involved? What is the structure and connectivity of the circuit? What are the signals flowing through the circuit? How do these circuit functions relate to behavior and cognition?" DOE laboratories clearly have expertise that relates to these goals ... "

As one of the first experimental neuroscientists at a DOE National Lab, I am amazed every day at the resources and tools that are at my disposal for brain science. The imaging technologies and advanced data-analysis techniques available through the Argonne Leadership Computing Facility (ALCF) and the Advanced Photon Source (APS) enable me to map the intricacies of brain function at the deepest levels and to describe these processes in greater detail than ever before. Those tools will be even more powerful in the future. The upgrade to the APS will create the ultimate 3-D microscope, producing the world's brightest hard x-rays and transforming our ability to understand and manipulate matter—including brains—at the nanoscale.

In 2021, Argonne will deploy the Aurora supercomputer at the ALCF. Aurora will be the first exascale-class system—at least 50 times faster than the nation's most powerful supercomputers in use today—in the United States. Aurora will enable us to explore new frontiers in artificial intelligence and machine learning; this will be the first time scientists have had a machine powerful enough to match the kind of computations the brain can do. It will be a breakthrough for neuroscience and for modeling biological processes. With the help of Aurora, I will be able to piece together millions of two-dimensional images, reconstructing the brain in three dimensions to create a map of the human brain.

These world-class user facilities—particularly when leveraged together—are and will continue to be critical to my efforts. For example, current recording and imaging methods can sample only a limited number of neurons or limited brain volumes, which constrains neuroscientific discovery. However, when data from imaging facilities like the APS is later modeled, simulated, and analyzed on a DOE supercomputer, neuroscientists can image and analyze every cell and blood vessel in a series of complete mammalian brains. Using one of the current fastest supercomputers on the planet at Argonne, called Mira, —I can quickly and efficiently analyze the millions of gigabytes of data this will produces. Imagine the game-changing possibilities of a resource where neuroscientists around the U.S., and ultimately around the world, utilize such technologies and infrastructure.

As members on the House Science, Space & Technology Committee, you understand that there are pivotal moments in science that we can harness to advance society in leaps, rather than small steps. Brain research is at that critical moment now. Neuroscientists are glimpsing a future where we can potentially understand the physical bases of mental illnesses that currently impose huge personal and financial burdens. Computer scientists see a future where the U.S. leads the world in computer science and artificial intelligence, which is critical for our national security and economic progress. The U.S. leads in both fields, for now, but we have not made brain mapping a national priority as Japan, South Korea, and China all have. The moment to cement our national leadership is now.

In 1962 at Rice University, President John F. Kennedy announced that the United States would put a man on the moon. Seven years later, Neil Armstrong walked on the face of the moon.

While some may say that the endeavor was a failure—where are moon bases now?—it is worth noting that when we landed on the moon in 1969, the average age of a NASA scientist was 29 years old. Seven years earlier, at the time of Kennedy's announcement, these scientists were college students seeking inspiration. The "moon shot" changed their lives and focused their passion so that they could change society in innumerable ways.

If we seize the opportunity now for a national moon shot for the brain, if we inspire the next generation of students to work at the intersection of brain science, computer science, and big data, we can make significant progress toward understanding the brain and curing brain diseases. We can create the next generation of computers and robots based on the brain, transforming our society and assuring U.S. leadership in these vital realms for the future. Thank you for your time and attention today. I welcome any questions you may have.

Dr. Kasthuri is the first Neuroscience Researcher at Argonne National Labs and an Assistant Professor in the Dept. of Neurobiology, University of Chicago. He has an MD from Washington University School of Medicine and a D.Phil. from Oxford University where he studied as a Rhodes scholar. As a post-doctoral fellow, Dr. Kasthuri developed an automated approach to large volume serial electron microscopy ('connectomics'). Currently, the Kasthuri lab continues to innovate new approaches to brain mapping including the use of high-energy x-rays from synchrotron sources for mapping brains in their entirety. The Kasthuri lab is applying these techniques to in service of answering the question: how do brains grow up, age, and degenerate?

Chairman WEBER. Thank you, Doctor.
Dr. Yelick, you're recognized for five minutes.

TESTIMONY OF DR. KATHERINE YELICK, ASSOCIATE LABORATORY DIRECTOR FOR COMPUTING SCIENCES, LAWRENCE BERKELEY NATIONAL LABORATORY; PROFESSOR, THE UNIVERSITY OF CALIFORNIA, BERKELEY

Dr. YELICK. Chairman Smith, Chairman Weber, Chairwoman Comstock, Ranking Members Veasey and Lipinski, distinguished Members of the Committee, thank you for holding this hearing and for the Committee's support for science. And thank you for inviting me to testify.

My name is Kathy Yelick and I'm the Associate Laboratory Director for Computing Sciences at Lawrence Berkeley National Laboratory, a DOE Office of Science laboratory managed by the University of California. I'm also Professor of Electrical Engineering and Computer Sciences at the University of California, Berkeley.

Berkeley Lab is home to five national scientific user facilities serving over 10,000 researchers covering all 50 States. The combination of experimental, computational, and networking facilities puts Berkeley Lab on the cutting edge of data-intensive science.

In my testimony today, I plan to do four things: first, describe some of the large-scale data challenges in the DOE Office of Science; second, examine the emerging role of machine learning; third, discuss some of the incredible opportunities for machine learning in science, which leverage DOE's role as a leader in high-performance computing, applied mathematics, experimental facilities, and team-based science; and fourth, explore some of the challenges of machine learning and data-intensive science.

Big-data challenges are often characterized by the four "V's," the volume, that is the total size of data; the velocity, the rate at which the data is being produced; variability, the diversity of different types of data; and veracity, the noise, errors, and the other quality issues in the data. Scientific data has all of these.

Genomic data, for example, has grown by over a factor of 1,000 in the last decade, but the most abundant form of life, microbes, are not well-understood. Microbes can fix nitrogen, break down biomass for fuels, or fight algal blooms. DOE's Joint Genome Institute has over 12 trillion bases—that is DNA characters A, C, T, and G—of microbial DNA, enough to fill the Library of Congress if you printed them in very boring books that only contain those four characters.

But genome sequencers produce only fragments with errors, and the DNA of the entire microbial community is all mixed together. So it's like taking the Library of Congress, shredding all of the books, throwing in some junk, and then asking somebody to reconstruct the books from them. We use supercomputers to do this, to assemble the pieces, to find the related genes, and to compare the communities.

DOE's innovations are actually helping to create some of these data challenges. The detectors used in electron microscopes, which were developed at Berkeley Lab and since commercialized, have produced data that's almost 10,000 times faster than just ten years ago.

Machine learning is an amazingly powerful strategy for analyzing data. Perhaps the most well-known example is identifying images such as cats on the internet. A machine-learning algorithm is fed a large set of, say, ten million images of which some of them are labeled as having cats, and the algorithm uses those images to build a model, sort of a probability of which images are likely to contain cats. Now, in science we're not looking for cats, but images arise in many different scientific disciplines from electron microscopes to light sources to telescopes.

Nobel laureate Saul Perlmutter used images of supernovae—exploding stars—to measure the accelerating expansion of the universe. The number of images produced each night from telescopes has grown from tens per night to tens of millions per night over the last 30 years. They used to be analyzed manually by scientific experts, and now, much of that work has been replaced by machine-learning algorithms. The upcoming LSST telescope will produce 15 terabytes of data every night. If you watch that, one night's worth of data as a movie, it would take over ten years, so you can imagine why scientists are interested in using machine learning to help them analyze that data.

Machine learning can be used to find patterns that cluster similar items or approximate complicated experiments. A recent survey at Berkeley lab found over 100 projects that are using some form of machine learning. They use it to track subatomic particles, analyze light source data, search for new materials for better batteries, improve crop yield, and identify abnormal behavior on the power grid.

Machine learning, it does not replace the need for high-performance computing simulations but adds a complementary tool for science. Recent earthquake simulations of the bay area show that just a 3-mile difference in location of an identical building makes a significant difference in the safety of that building. It really is all about location, location, location. And the team that did this work is looking at taking data from embedded sensors and eventually even from smart meters to give even more detailed location-specific results.

There is tremendous enthusiasm for machine learning in science but some cautionary notes as well. Machine-learning results are often lacking in explanations, interpretations, or error bars, a frustration for scientists. And scientific data is complicated and often incomplete. The algorithms are known to be biased by the data that they see. A self-driving car may not recognize voices from Texas if it's only seen data from the Midwest.

Chairman WEBER. Hey, hey.

Dr. YELICK. Or we may miss a cosmic event in the southern hemisphere if they've only seen data from telescopes in the northern hemisphere. Foundational research in machine learning is needed, along with the network to move the data to the computers and share it with the community and make it as easy to search for scientific data as it is to find a used car online.

Machine learning has revolutionized the field of artificial intelligence and it requires three things: large amounts of data, fast computers, and good algorithms. DOE has all of these. Scientific instruments are the eyes, ears, and hands of science, but unlike artificial intelligence, the goal is not to replicate human behavior but to augment it with superhuman measurement control and analysis capabilities, empowering scientists to handle data at unprecedented scales, provide new scientific insights, and solve important societal challenges.

Thank you.

[The prepared statement of Dr. Yelick follows:]

Big Data Challenges and Advanced Computing Solutions

A Hearing of the

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
UNITED STATES HOUSE OF REPRESENTATIVES

Testimony of

DR. KATHY YELICK, ASSOCIATE LABORATORY DIRECTOR FOR COMPUTING SCIENCES
LAWRENCE BERKELEY NATIONAL LABORATORY

July 12, 2018
2318 Rayburn House Office Building
Washington, DC

INTRODUCTION

Chairman Weber, Chairwoman Comstock, Ranking Members Veasey and Lipinski, and distinguished Members of the Committee, thank you for holding this hearing and for the Committee's support for science. The opportunities presented by "Big Data" are advancing science and innovation in novel and exciting ways. Machine learning is an important part of this story and I commend the committee for exploring how new capabilities in high performance computing and computational science will open doors to new knowledge.

My name is Kathy Yelick and I am the Associate Laboratory Director for Computing Sciences at Lawrence Berkeley National Laboratory, a DOE Office of Science laboratory managed by the University of California. I am also a Professor of Electrical Engineering and Computer Sciences at the University of California, Berkeley. It is my honor and my pleasure to participate in this hearing and to aid the Committee's examination of the opportunities and challenges related to data analytics and machine learning within the Department of Energy. Thank you for inviting me to testify.

Berkeley Lab is a multipurpose lab with world leading capabilities across materials research, biosciences, physics, chemical sciences, energy technologies, earth and environmental sciences, high performance computing, advanced networking and more. Home to five national scientific user facilities, Berkeley Lab serves over 10,000 researchers from all 50 states and beyond. Thirteen Nobel prizes are associated with Berkeley Lab, as are fifteen National Medal of Science recipients. Seventy Berkeley Lab scientists are members of the National Academy of Sciences, one of the highest honors for a scientist in the United States, eighteen of our engineers have been elected to the National Academy of Engineering, and three of our scientists have been elected into the National Academy of Medicine. In addition, Berkeley Lab has trained thousands of university science and engineering students who are advancing technological innovations across the nation and around the world.

In my testimony today I plan to do four things:

First, describe some of the large scale data challenges in the DOE Office of Science, drawing examples from Berkeley Lab and other national laboratories' national user facilities and team science projects.

Second, talk about the emerging role of machine learning - and specifically deep learning - methods, which have revolutionized the field of artificial intelligence (AI) and may similarly impact scientific discovery.

Third, discuss some of the unique opportunities for machine learning in science, leveraging DOE's national role as a leader in high performance computing, applied mathematics, user facilities, and interdisciplinary team science.

And, **fourth,** describe a vision for the national laboratories that includes foundational research in data science along with an interconnected network of experimental and computational facilities to address some of the most challenging data analytics problems in science.

Part 1: Data challenges in science
The Department of Energy has a unique role in science as the largest funder of physical sciences research in the nation and with the responsibility for managing and operating many of the largest scientific user facilities. At Berkeley Lab alone, as mentioned previously, there are over 10,000 users of our scientific user facilities, which include the Advanced Light Source, the Joint Genome Institute, the Molecular Foundry, National Energy Research Scientific Computing Center (NERSC), and the Energy Sciences Network (ESnet). In addition, the Lab is a partner in many national and international collaborations, such as the ATLAS, Alice, and CMS projects at the Large Hadron Collider (LHC) in Switzerland, the Dark Energy Spectroscopic Instrument (DESI) near Tucson, Arizona, and the LZ dark matter experiment in South Dakota.

Big data challenges are often characterized by the *4 Vs*: volume (the total size), velocity (the speed at which is it being produced), variability (the diversity of data types) and veracity (noise, errors, and other quality issues). Scientific data has all of these, and DOE's user facilities are a big source of the challenges and the opportunities to use large data sets for new discoveries, because of increasing data rates, reduced costs of collecting data, and total data volumes.

The cost of sequencing the human genome is now around $1,000 down from $10,000,000 just a decade ago, and the National Institutes of Health (NIH) database on genomic data (the Sequence Read Archive, or SRA) now holds over 8 petabytes (10^{15} bytes) of genomic data, a 3000x increase in ten years. At DOE's Joint Genome Institute, a newer database of viral genomes has grown nearly 100x in just two years.

In cosmology and particle physics, the velocities and volumes have also grown, with the upcoming Large Synoptic Survey Telescope (LSST) producing about 20 terabytes (10^{12} bytes) every night and a resulting community data set over its lifetime of about 60 petabytes. The LHC will collect roughly 50 petabytes of data in 2018, even after eliminating 99% of the data produced inside the experiment, and that 50 petabytes will grow to roughly 500 petabytes by 2024. The LHC data from past experiments is copied to data centers around the world with 900 petabytes currently in disk and tape storage.

The volume and velocity of scientific data is growing because the instruments are improving -- we can see things at a microscopic and atomic scale, measure vibrations imperceptible to the human eye, and take high resolution images of objects in the universe that are millions of light years away. The national labs are key to developing many of the instruments used for major science experiments. For example, Berkeley Lab has a long history of developing the detectors used in electron and x-ray microscopy, improving spatial resolution 100-fold and temporal resolution 1,000-fold, to reveal atomic structure without the need for crystallization. This technology was revolutionary in chemistry, material science, and biology, and its use in Cryo Electron Microscopy instruments was cited in the 2017 Nobel Prize in Chemistry, as well as being commercialized for advanced medical and scientific imaging.

Data veracity and variety are also challenges in nearly every discipline in science, with scientists eager to extract vanishingly small signals from large messy data sets and combine different modalities to improve insights. One of the most exciting fields for the application of big data science and advanced computing is biology - the increasing sizes of data sets, the inherent noise, and the complexity make it a prime area of research to leverage these emerging tools and capabilities.

Microbes are a data challenge - they are the most abundant and diverse life form on Earth. They exist in vast complex microbial communities called microbiomes and interact with and significantly impact all of the world's natural systems, including human, plant, animal, energy, and environmental, at all scales, from the infinitesimal to the grand. In one handful of soil there may be as many microbes as there are stars in our galaxy. Discovering how these complex communities of millions and billions of microbes interact and impact natural systems will create new knowledge and advance solutions to the world's most intractable problems - unlocking and harnessing the mysteries of microbiomes will advance environmental remediation, propel new agricultural technologies and processes, and speed biologically-based energy solutions to market. This research will grow the United States' bioeconomy and drive tremendous economic activity. Maintaining U.S. leadership in microbiome research is an economic and national security strategic imperative - but, it's a hard nut to crack, in large part due to the data challenges.

For example, a particular microbial species and those its genetic data often occur only in samples with hundreds of different species and thousands of strains mixed together. To further complicate analysis, today's standard sequencing technologies produce error-laden fragments of DNA that need to be assembled together to find genes. Putting it all together in a scientifically useful way is analogous to completing hundreds of different jigsaw puzzles with all the pieces mixed together, some of the pieces broken, and no reference pictures for any of the final images. Of course, the function of a microbiome is more than the sum of the parts, with multiple species interacting to impact the environment in which they live, whether it's within the human body or in the environment. To understand and eventually control microbial behavior from the genomic level, one also must combine genomic data with a variety of data from imaging, chemical sensors, and other scientific instruments - making an already complex task more difficult.

High performance computing and novel computational methods give scientists the tools needed to decipher these microbial puzzles and to assemble, shape and coax the data into useable information, removing errors, finding genes, and discovering relationships between species and across different microbiome samples. My own research includes leading the microbiome application project in the Exascale Computing Project (ECP), where the overarching goal is to find new information about the microbiome using more powerful algorithms and computers. It may reveal changes in the microbiome due to diet, weather, chemicals or other environmental factors, and ways of using a microbial community to produce a healthier environment for humans as well for food crops. More broadly, scientists have recognized the need for interdisciplinary research efforts in this area and for a National Microbiome Data Collaborative, an open, standardized, and shared data infrastructure, that could help foster integrated analyses and synthesis across diverse microbial datasets.

As another example of an ECP application, scientists and engineers at Berkeley Lab and Lawrence Livermore National Laboratory (LLNL) recently performed a simulation of a large-magnitude earthquake in the San Francisco Bay Area, and how it would affect different locations and buildings, with the goal of understanding impacts on critical infrastructure such as schools, hospitals, and the power grid. This was done at an unprecedented scale and resolution using Berkeley Lab's NERSC supercomputers. Even larger simulations will be done on future exascale systems. Already, these simulations have shown that the same building located less than 3 miles apart may have different risks and therefore require different building hardening.

To make the results even more specific to a given location, the team is looking at using measured seismic data obtained from regionally deployed sensors during frequently occurring small earthquakes to help improve fine-scale geologic models. By computationally "inverting" measured data, enhanced understanding of the subsurface geology can be obtained to improve the computational models for ground motion simulations. Merging high performance simulations with measured data will yield even more precise information about shaking at every location throughout the region. While this massive aggregation is no small undertaking, it will lead to improved public safety. These types of approaches are only becoming feasible because of the major advancements being made in high performance simulation combined with big data exploitation.

Looking towards the future, even more dense data will become available as seismic sensors proliferate. For example, recent technology advancements provide an opportunity to deploy seismic sensors across the electric grid onboard smartmeters, which will provide unprecedented data. We need to be prepared to exploit this emerging big data for transforming hazard and risk assessments.

Part 2: Data Analytics, Machine Learning, Deep Learning, and Artificial Intelligence

Machine learning is an increasingly popular strategy for analyzing scientific data and offers opportunities to better leverage and benefit from the explosion in data volume. It's also a term that is used very broadly to refer to methods that learn from data, or to make inferences based on a model learned from some data. The most well-known example is identifying images, such as cats, on the internet. A machine learning algorithm is fed a large set of, say, 10 million images of which some are labeled as having a picture of a cat. The algorithm uses those examples to build a model of which images contain cats, i.e., the probability that a given image contains a cat. For example, an image with two diagonal lines that meet to form something like a cat's ear will have a higher probability of containing a cat. This is known as *supervised learning*, because one starts with images labeled as cats or not, whereas *unsupervised learning* might have a set of unlabeled images and can determine which ones have similar objects in them, but does not determine what the objects are. In science we are not looking for cats, but we can use machine learning to find features such as exploding stars, cellular structures, or subatomic particles.

There are several different kinds of machine learning algorithms, but many of the most notable breakthroughs in recent years have come from a powerful class of *deep learning* algorithms, and specifically *deep neural networks*, which are used in this example of finding cats or other images in internet searches. The algorithm works in a set of layers, where one layer may find the edges between different objects in a picture, another layer will find shapes from the edges, and higher level layers will find recognizable objects such as eyes, ears, and tails. The number of layers will vary depending on the application problem, and it is one of the things that a data scientist may have to experiment with, but typically the depth is a few layers to a few dozen.

Deep learning has led to a number of surprising results in the field of Artificial Intelligence (or AI), which has the goal of developing computers with human-like capabilities, including computer vision, speech recognition, robotics, and playing games of strategy. Deep learning is used in Siri to recognize speech and interpret commands, and it is used in self-driving cars to identify road signs, hazards, and obstacles. It was also used by Alibaba, a Chinese company, to outperform students on a standardized reading comprehension exam, similar to what is used in college admissions, and by Google's AlphaGo in 2016 to beat the world ranking world champion (Lee Sedol) in the game of Go, a strategy board game that is significantly harder than chess. These deep learning methods are so strongly linked with AI that the terms *AI* and *deep learning* are sometimes used synonymously.

These ideas have been around for a long time -- the neural net ideas go back to the 1940s, and the key algorithmic idea (*backpropagation*) was developed in the 1980s. So, why have these algorithms suddenly become successful? Roughly speaking, machine learning requires three things: large amounts of data, fast computers, and good algorithms. The growth in data has been fueled by the ubiquity of cameras, recording devices, and various sensors, facilitated by the ease of sharing data over the Internet, while computing performance has grown by a factor of one million since the early 80s. As described earlier, there has been a similar explosion in data in science coming from instruments that provide more detail and higher data rates, and from increasingly complex simulations enabled by faster computers. With DOE's abundant use of simulation, faster computers are both part of the challenge and part of the solution.

DOE's unique resources in high end computing have also proven to be well suited to machine learning, and deep learning maps well onto the Graphics Processing Units (GPU) in the pre-exascale systems recently deployed in the Summit machine at the Oak Ridge Leadership Computing Facility (OLCF) and Sierra at LLNL. One of the key computational kernels in deep learning is multiplying two matrices, which also is the dominant kernel in the Linpack benchmark used for the TOP500 list, where Summit and Sierra are in the #1 and #3 spots respectively. Not surprisingly, some of the early science projects on these computers are focused on machine learning, and are using the high speed networks, large amounts of memory, and unprecedented computing capability to solve problems that are intractable on conventional computers.

Each year the top performing scientific application team is awarded the Gordon Bell prize, an award that reflects science at scale, as opposed to a fixed benchmark. Finalists for the 2018 prize include a deep learning computation at over 200 petaops[1]/sec computation on Summit, which was a partnership between NERSC, OLCF, NVIDIA, and Google, that was used to analyze data from cosmology and extreme weather events. A second finalist is a project lead by Oak Ridge National Laboratory with researchers from the University of Missouri in St. Louis, which used an entirely different algorithm to learn relationships between genetic mutations across an enormous set of genomes, with potential applications in biomanufacturing and human health. This algorithm can also be mapped to matrix-multiply like operations. It runs at a impressive 1.88 exaop/second! These are the fastest deep learning and other machine learning computations to date.

Part 3: The use of machine learning in science

What does this mean for science? The image analysis example is directly analogous to science, because images arise in many scientific disciplines, from electron microscopes in biology, to x-rays from light sources used in material science, to telescopes used in cosmology. Saul Perlmutter, a Nobel Laureate from Berkeley Lab, used image analysis to discover the accelerating expansion of the Universe through observations of distant supernovae - exploding stars - as a kind of standard reference point. His team used a specific kind of supernova, Type 1A, which occurs when a white dwarf star explodes; these are fairly rare, with about one per century within our Milky Way Galaxy. Using high powered telescopes and collecting images over many months, they would look for the appearance of new stars in remote galaxies that suddenly appeared on an image (called a "transient") and then use other telescopes, like the Hubble Space Telescope and the Keck Telescopes, to classify the transient as a variable star, a quasar, or supernova.

Thirty years ago, a few tens of images were produced each night and were analyzed manually by scientific experts. By 2007, some automatic processing was done to find transients, and Berkeley Lab was already working on using machine learning algorithms to classify supernovae. Today, tens of millions of images are produced from experiments like the Dark Energy Survey, the Zwicky Transient Facility, and soon the Large Synoptic Survey Telescope, which produce thousands of new transient discoveries each night. Machine learning algorithms run automatically on supercomputers at NERSC and the National Science Foundation's National Center for Supercomputing Applications center in Illinois, scouring these images each night for new transients. These machine learning algorithms make scientists much more productive, reducing by more than 10,000x the number of images they look at manually. Today, the focus is on using deep learning to not only find these transients, but to classify them so that follow-up resources are only spent on those objects the scientists want to study.

---

[1] Petaop/second and exaop/second are, respectively, 10^{15} and 10^{18} operations per second. For machine learning applications, the computations can often be performed using less powerful operations than normal (double precision) floating point operations (*flops*) used in High Performance Computing (HPC). These machine learning "ops" are about one quarter as powerful those typical HPC flops.

Machine learning can often find patterns in noisy data when other approaches fail, so scientific applications are not limited to cosmology or even to images. We recently surveyed researchers at Berkeley Lab and found over 100 projects, many in partnership with other labs and universities, that are using some form of machine learning. For example, researchers from Fermilab, Caltech, and Berkeley Lab are exploring the use of deep learning to identify and track particles in experiments at the Large Hadron Collider, working to replace current algorithms that are not easily implemented on high performance computers using traditional approaches, and are projected to consume enormous amounts of computing time after the LHC upgrade. Another group has developed machine learning strategies that aim to increase crop yields and improve the sustainability of agriculture while reducing economic risks for farmers and landowners as part of the AR1K collaboration between Berkeley Lab, the University of Arkansas, and Glennoe Farms. The farm is an experimental platform instrumented with sensors, drone-based imaging, and frequent data collection, and machine learning is the tool that will tie all the data together. The Lab's Center for Advanced Mathematics for Energy Research Applications (CAMERA) has developed a new deep learning algorithm to analyze light source images more quickly and more accurately than previous approaches, and the Materials Project is using machine learning to remove the guesswork from materials discovery and design, driving the development of advanced materials.

At DOE's Joint BioEnergy Institute (JBEI), scientists are using machine learning to improve biosensor design and accelerate synthetic biology to produce biofuels; the technique can predict the amount of biofuel produced by newly engineered bacterial cells based on data from previous experiments and has other applications, such as developing drugs that fight antibiotic-resistant infections and crops that withstand drought. And at the Joint Genome Institute, machine learning is used to answer fundamental questions about biology, such as the relationships between all genes that naturally occur in the environment.

DOE researchers are also using machine learning to improve the operation of its facilities and make scientists more productive. An enormous challenge in large data is getting labels or metadata information associated with data from each scientific experiment - making it more accessible and useful to scientists. Researchers at Berkeley Lab have developed techniques to automatically label data, starting with the enormous stream of data on advanced materials for batteries and other applications, coming from the National Center for Electron Microscopy (NCEM), home to the world's most powerful electron microscope. In another example, ESnet, DOE's advanced high speed scientific network, is using a variety of machine learning techniques to predict the amount of data being transferred between endpoints in order to adapt network traffic dynamically, detect problems in the infrastructures, and find anomalous high-volume data transfers, which could indicate either a faulty device or a cyber attack.

Similar ideas are used for other real-time flows of time-series data, such as looking for abnormal behavior in the power grid, in rooftop solar panel systems, or even financial market data. Data from cell phones and embedded sensors are being used to build large-scale models of regional transportation systems, which can be used for long term planning of transportation infrastructure, energy planning, and emergency response.

Machine learning expertise at the DOE labs can also be leveraged for other national priorities, as in partnerships with the NIH and the Department of Veterans Affairs (VA). The data in this case includes genomes, images, results of medical tests, and electronic medical records. It is being used to address medical challenges such as traumatic brain injury, cancer, mental health, and the opioid crisis.

Berkeley Lab researchers have developed machine learning approaches to analyze and visualize brain image data collected from multiple devices, to automatically identify cancer cells in image data, fibers in textile images, and more. DOE researchers bring expertise on high performance computing, data analytics, modeling and simulation, as well as a culture of team-based science, where the team of cross-disciplinary scientists and engineers work on end-to-end solutions for grand challenge problems.

Machine learning does not replace the need for the more traditional use of HPC simulations, but instead offers a complementary set of techniques. Roughly speaking, simulation is used when a set of equations, i.e. a model, is known in advance, while machine learning may be used to infer a model based on data from observations. Machine learning is often combined with simulation to fill in parts of simulation where no known equations exist but where data is available. This approach is being used for simulating turbulent fluids by researchers at Sandia.

Machine learning is also used to accelerate large ensembles of simulations, where the machine learning can quickly approximate them to determine which ones are most important in searching for a particular outcome. Finally, machine learning can be performed on the data produced by simulations, such as in research at Berkeley Lab searching for extreme weather events. As stated in a recent report by the DOE Advanced Scientific Computing Research (ASCR) community on scientific machine learning, "In all cases, it is clear that ML will not replace decades of research in principled physics-based approaches. Rather, it can bring a toolbox of methods to enrich, improve, and accelerate current modeling approaches."

Part 4: The need for foundational research and interconnected facilities to advance data-driven scientific discovery

As indicated from the examples above and many more across the national lab complex, scientists are actively pursuing the use of machine learning and advanced analytics in nearly every basic and applied scientific domain. Enthusiasm is appropriate based on existing results from AI and from the success of many commercial applications. However, enthusiasm should be tempered with some understanding of the challenges facing scientific applications. ASCR's recent workshop on scientific machine learning elucidated many of these challenges and the need for additional research on the mathematical foundations of machine learning, including the following topics:

1) Leveraging scientific domain knowledge: Many machine learning techniques have been developed primarily for images, speech, and textual data. Speech and textual data may have use in analyzing scientific publications, notebooks, and presentations, but would not be the

primary focus. And, while images are common in science, their formats and content will be quite different than in more common internet image searches. Much of scientific data from simulations is three dimensional and may exhibit symmetries not recognized by current algorithms, e.g., in molecular structures. Finally, scientific data is often incomplete and acquiring large sets of labeled data for supervised learning may not be practical.

) Interpretable machine learning: The models derived from machine learning methods, and especially deep learning, are not a recognizable set of equations but instead may involve many thousands of parameters without an explanation of what each one means. For placing advertisements or classifying images, this may be acceptable, as long as they make good choices, but scientists will demand more confidence, better error bars, and in some cases direct explanations for use in developing theories. Simple correlations are not sufficient.

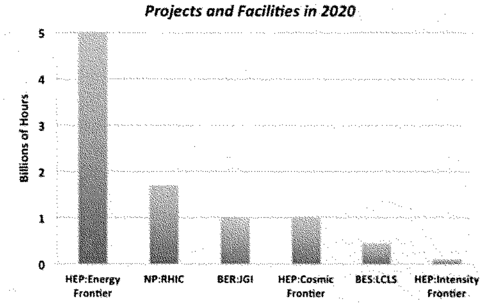
) Robust and efficient: There should be rigorous numerical estimates for the quality of results and well-defined criteria on when the inferences may be used. When applying machine learning to infer properties of data, one distinguishes between that data given to the algorithm for "training" and the data on which it will be used. For example, images labeled as containing a supernova may be given as training, and once trained, it can be used to automatically label images. Methods should be insensitive to the details of how training data is selected and should perform well as long as the training set is in some sense reasonable. In AI research, there are concerns about bias that derives from the selection of training data -- only posting advertisements for a CEO position to white men, for example. In science, bias may come from artifacts of a particular instrument or measurement technique.

The ASCR workshop report also recognized the close interaction of simulation and machine learning, as well as opportunities to control large scientific campaigns, whether a large set simulations or experiments, to choose the best cases to run.

While the workshop was targeting mathematical research in machine learning, there are also computer science problems related to programmability, performance, parallelization, and scale. There is currently a strong preference in the deep learning community for GPU architectures, but even more specialized architectures may prove beneficial, which could create a divergence between the computing platforms for simulation and learning. Other machine learning algorithms may place higher demands on the memory system and network, and as the foundational work evolves, deep learning methods may use sparsity, e.g., to improve running time or interpretability, which will also shift hardware requirements.

DOE will need to address the burgeoning data analytics needs from major experiments and embedded sensors, whether those use deep learning, other machine learning methods, or other analytics techniques. The figure on the right shows the

Large Compute Requirements from DOE Experimental Projects and Facilities in 2020



estimated computing requirements of some of the major DOE experiments, normalized to NERSC computing hours. For comparison, NERSC currently delivers about seven billion hours to its users. Some of the data analysis will require real-time processing, automated job scheduling, and other policy changes, in addition to sufficient computing to meet this growing scientific demand.

Conclusions

Data-driven scientific discovery is poised to deliver breakthroughs across many disciplines, and as stewards of many national user facilities, DOE should have a leadership role. Driven by innovations in instrumentation and computing, and a desire to investigate increasingly complex scientific questions, the data challenges will continue to grow. In addition to analytics problems, there are many technical challenges in data curation, sharing, metadata labeling, and search, to give scientists tools for research that are analogous to those that have revolutionized shopping, entertainment, and business.

Machine learning is a promising approach for analytics in science, complementing but not replacing modeling and simulation. In spite of the extensive work already going on at the DOE labs, machine learning and associated mathematical foundations of machine learning are not as well developed as in simulation science.

The goal in scientific discovery is more focused than so-called *general AI*, but also goes beyond emulating human capabilities. The scientific instruments described earlier are the eyes, ears and hands of science, and the goal is not to replicate human behavior but augment it with superhuman measurement, control and analysis capabilities, empowering scientists to ask and answer more complex questions. For this reason the alternate phrase, *Intelligence Augmentation (IA),* is probably more appropriate.

The Exascale Computing Initiative is addressing one of the three requirement to make machine learning successful in DOE: availability of extreme computing capabilities. The Exascale Computing Project is addressing some of the underlying computational challenges of data analytics, with applications in cancer research, microbiome analysis, and light source imaging, all involving some form of machine learning along with other simulation and analytics methods. The project also has co-design centers in graph analytics and machine learning, which are linked to a number of the 24 applications. Along with the procurement strategies at the computing facilities, ECP will ensure that future exascale system architectures are well suited to this workload. But the foundational research in machine learning and broader facility issues still needs to be addressed.

While raw data is growing, DOE will need a strategy and infrastructure to enable sharing, search, curation and management, and to ensure the facilities are coupled in a way that large experiments can take advantage of the high end computing facilities for the largest data challenges

In the excitement over machine learning methods and data produced by new instruments and embedded sensors, one should not lose sight of the need for stronger foundational work. DOE has taken this seriously in the field of modeling and simulation, developing mathematical models and algorithms with proven performance and quality metrics, along with quantifiable measures of uncertainty and errors. The scientific peer review process will drive machine learning to be similarly rigorous, in a way that the commercial applications do not. DOE's interdisciplinary approach, which will require additional expertise in statistics and mathematical optimization, in addition to current strengths in applied mathematics and computer science, should lead to high quality methods that solve real problems and lead to new methods and insights that will benefit other applications of machine learning.

Katherine (Kathy) Yelick is a Professor of Electrical Engineering and Computer Sciences at UC Berkeley and the Associate Laboratory Director (ALD) for Computing Sciences at Lawrence Berkeley National Laboratory. Her research is in high performance computing, programming languages, compilers, parallel algorithms, and automatic performance tuning. She currently leads the Berkeley UPC project and co-lead the Berkeley Benchmarking and Optimization (Bebop) group. As ALD for Computing Sciences at LBNL, she oversees the National Energy Research Scientific Computing Center (NERSC), the Energy Sciences Network (ESnet) and the Computational Research Division (CRD), which covers applied math, computer science, data science and computational science.

Chairman WEBER. Thank you, Doctor.
Dr. Nielsen, you're recognized for five minutes.

TESTIMONY OF DR. MATTHEW NIELSEN, PRINCIPAL SCIENTIST, INDUSTRIAL OUTCOMES OPTIMIZATION, GE GLOBAL RESEARCH

Dr. NIELSEN. Chairman Smith, Chairman Weber, and Chairwoman Comstock, Ranking Members Veasey and Lipinski, and Members of the Subcommittee, it is an honor to share General Electric's perspective on innovative machine-learning-based approaches to big-data science challenges that promote a more resilient, efficient, and sustainable energy infrastructure. I am Matt Nielsen, a Principal Scientist at GE's Global Research Center in upstate New York.

The installed asset base of GE's power and renewable businesses generates roughly 1/3 of the planet's power, and 40 percent of the world's electricity is managed by our software. GE Energy's assets include everything from gas and steam power, nuclear, grid solutions, energy storage, onshore and offshore wind, and hydropower.

The nexus of physical and digital technologies is revolutionizing what industrial assets can do and how they are managed. One of the single most important questions industrial companies such as GE are grappling with is how to most effectively integrate the use of AI and machine learning into their business operations to differentiate the products and services they offer. GE has been on this journey for more than a decade.

A key learning for us—and I can attest to this as being a physicist—has been the importance of tying our digital solutions to the physics of our machines and to the extensive knowledge on how they are controlled. I'll now highlight a few industrial applications of AI machine learning where GE is collaborating with our customers and federal agencies like the U.S. Department of Energy.

At GE, digital twins are a chief application of AI and machine learning. Digital twins are living digital models of industrial assets, processes, and systems that use machine learning to see, think, and act on big data. Digital twins learn from a variety of sources, including sensor data from the physical machines or processes, fleet data, and industrial-domain expertise. These computer models continuously update as new data becomes available, enabling a near-real-time view of the condition of the asset.

To date, GE scientists and engineers have created nearly 1.2 million digital twins. Many of the digital twins are created using machine-learning techniques such as neural networks. The application of digital twins in the energy sector is enabling GE to revolutionize the operation and maintenance of our assets and to drive new innovative approaches in critical areas such as services and cybersecurity.

Now onto digital ghosts. Cyber threats to industrial control systems that manage our critical infrastructure such as power plants are growing at an alarming rate. GE is working with the Department of Energy on a cost-shared program to build the world's first industrial immune system for electric power plants. It cannot only detect and localize cyber threats but also automatically act to neutralize them, allowing the system to continue to operate safely.

This effort engages a cross disciplinary team of engineers from the global research and our power business. They are pairing the digital twins that I mentioned of the power plants machines, industrial controls knowledge, and machine learning. The key again for this industrial immune system is the combination of advanced machine learning with a deep understanding of the machines' thermodynamics and physics.

We have demonstrated to date the ability to rapidly and accurately detect and even localize simulated cyber threats with nearly 99 percent accuracy using our digital ghost techniques. We're also making significant progress now in automatically neutralizing these threats. It is a great example of how public-private research partnerships can advance technically risky but universally needed technologies.

Along with improving cyber resiliency, AI and machine-learning technologies are enabling us to improve GE's energy services portfolio, helping our customers optimize and reduce unplanned downtime for their assets. Through GE's asset performance management platform, we help our customers avoid disruptions by providing deep, real-time data insights on the condition and operation of their assets. Using AI, machine learning, and digital twins, we can better predict when critical assets require repair or have a physical fault. This allows our customers to move from a schedule-based maintenance system to a condition-based maintenance system.

The examples I have shared and GE's extensive developments with AI and machine learning have given us a first-hand experience into what it takes to successfully apply these technologies into our Nation's energy infrastructure. My full recommendations are in my written testimony, and I'll only summarize them here.

Number one, continue to fund opportunities for public-private partnerships to expand the application and benefits of AI and machine learning across the energy sector.

Two, encourage the collaboration between AI, machine learning, and subject matter experts, engineers, and scientists.

And number three, continue to invest in the Nation's high-performance computing assets and expand opportunities for private industry to work with the national labs.

I appreciate the opportunity to offer our perspective on how the development of AI and machine-learning technologies can meet the shared goals of creating a more efficient and resilient energy infrastructure.

One final thought is to reinforce a theme that I've emphasized throughout my testimony, and that is the importance of having teams of physical and digital experts involved in driving the future of AI and machine-learning solutions.

Thank you, and I look forward to answering any questions.

[The prepared statement of Dr. Nielsen follows:]

Testimony before
The U.S. House of Representatives Committee on Science, Space, and Technology,
Subcommittee on Energy and Subcommittee on Research and Technology
Thursday, July 12, 2018

Matthew Nielsen, Ph.D. – Principal Scientist, GE Global Research
Email: nielsema@ge.com

Chairman Weber and Chairwoman Comstock, Ranking Members Veasey and Lipinski, Representative Tonko and members of the Committee, it is an honor to share GE's perspective on innovative machine learning-based approaches to big data science challenges to promote a more resilient, efficient and sustainable energy infrastructure. I am Matt Nielsen, a Principal Scientist at GE's Global Research Center in Upstate New York specializing in the application of these digital technologies for GE's Power, Renewables and Aviation businesses.

Between GE's Power and Renewables businesses, GE has a $44 billion energy portfolio that powers 1/3 of the planet. And 40% of the world's energy is managed by our software. GE's energy assets include everything from gas and steam power, nuclear, grid solutions and energy storage to onshore and offshore wind and hydro power. Our application of advanced technologies has allowed us to achieve the world record in combined cycle gas turbine efficiency and recently introduce the world's largest offshore wind turbine, the 150M-6MW Haliade X.

GE Global Research was the first industrial research lab established in the United States in 1900 and today remains the cornerstone of innovation for the General Electric Company. We are home to one of the world's most diversified, interdisciplinary research organizations, with ~1,000+ scientists and engineers (~600 hold PhDs.) working at our research campus in Niskayuna, NY. It is at Global Research where GE's research activities in artificial intelligence (AI) and machine learning are being led to support our business interests and to unleash these technologies to help solve the world's toughest problems.

The nexus of physical and digital technologies is revolutionizing what industrial assets can do and how they are managed. One of the single most important issues industrial companies are grappling with is how to most effectively integrate the use of AI and machine learning into their business operations to differentiate the products and services they offer. GE has been on this journey for more than a decade, recognizing early on the impact digital technologies could have in taking our products and services to the next level of efficiency and performance.

A key learning for us has been the importance of tying our digital solutions to the physics of our machines and to our extensive knowledge on how they are controlled. To work in the industrial world, AI and machine learning technologies must be coupled with the laws of physics, or known truths about machines and the environment in which they operate.

My testimony focuses on a few industrial applications of AI and machine learning that GE is driving with our customers and with federal agencies like the U.S. Department of Energy to address key challenges with cybersecurity related to critical power assets and to enable a new services paradigm that minimizes and strives to eliminate unexpected disruptions in the operation of power generation assets. We call it zero unplanned downtime. Both examples would not be possible without this unique combination of physical and digital technologies.

The Industrial Internet of Things (IIoT)

The digitization of major industrial sectors including energy, aviation, transportation and healthcare represent the next frontier of the digital revolution we already have experienced in finance, entertainment and telecommunications. This current wave is being powered by the exponential growth in computing power and digital technologies that is allowing us to connect and control billions of machines. To illustrate this leap in technology, just consider that a typical gaming system you can buy for your kids for a few hundred dollars packs the same processing power as a $10 MM supercomputer from the late 1990s.

At GE, one of the chief manifestations of AI and machine learning is happening through GE's Digital Twin technology. Digital twins are living, digital models of industrial assets, processes and systems that use AI and machine learning technologies to see, think and act on big data to drive higher business value and outcomes for GE and our customers. GE's Twins learn from a variety of sources that include sensor data from the physical machines, systems or processes themselves, fleet data and industrial domain expertise from human engineers. These models continuously learn as new data comes in from one or more of these sources, enabling a real-time view of the condition of your assets at any point in time.

To date, GE scientists and engineers have created ~1.2 million Digital Twins of our industrial components, assets and processes that represent a broad cross-section of the energy, aviation, transportation and health care sectors. GE's Digital Twin is the platform for product lifecycle management from inception and design through operations and maintenance and all the way to decommissioning and repurposing of assets. In power specifically, the application of Digital Twins is enabling GE to revolutionize the maintenance and operation of our assets and drive new, innovative approaches in critical areas such as cybersecurity.

GE's Digital Ghost **– Building the World's 1st Industrial Immune System**

Cyber threats to Industrial control systems that manage critical infrastructure such as power plants are growing at an alarming rate. Between 2015 and 2016, the number of cyberattacks increased by 110%[1]. While we continue to see advances in IT and OT technologies to prevent attacks from getting through, GE is working with the Department of Energy (DOE) on a $4.1 million cost-shared program to build the world's first industrial immune system for electric power plants that could not only detect and localize cyber threats but automatically act to neutralize them, allowing the system to continue operating safely and efficiently.

[1] https://securityintelligence.com/attacks-targeting-industrial-control-systems-ics-up-110-percent/

The creation of an industrial immune system that monitors a power plant's assets 24/7 seeks to replicate the human body's response of automatically detecting and acting to stop a virus invading the body. To replicate this effect with industrial systems, a team of cross disciplinary engineers from GE Global Research and GE's Power business are pairing a complete Digital Twin of the power plant system (embedded with AI and machine learning technologies) with Industrial controls to trigger an automatic detection and response to cyber threats when a power system is under attack.

The premise for this industrial immune system to work is to understand the machine's physics. Around the Research Center we like to say, "you can't fool the physics." In other words, the Digital Twins, or digital models we create must mimic the actual physics of the power plant system itself. Fortunately for GE Global Research, we have both the digital and physical experts to design such a system.

Using sensors and controls, we're creating an immune system that will rapidly be able to detect and localize where a cyber threat is occurring using advanced AI techniques. But then, our cyber protection system will enable the power plant system itself to automatically respond by neutralizing the effects of the threat. Of course, we want the detection and response to cyber threats to be as fast as possible. This is another way we're using AI and machine learning. We can construct "reduced order" models, or mathematical models that can be executed very fast and quickly identify what's happening in the system. This allows fast detection of anomalies and the generation of rapid decisions on optimal settings to protect assets.

This simply can't be done without combining a deep knowledge of both the physics of the system with an extensive knowledge of industrial controls. Industrial controls are the brains of machines that control how they operate. Taking optimal control actions depends on getting the best data insights through AI and machine learning technologies.

The program with the DOE is ongoing. To date, we have demonstrated the ability to rapidly and very accurately detect (99% accuracy) and localize simulated threats. We are also making progress on designing a system that can automatically act to neutralize threats. It is a great example of the tremendous value public/private research partnerships can advance technically risky, universally needed technologies that stay a step ahead of cyber attackers and strengthen the resilience of our energy infrastructure.

Uninterrupted Power... Zero Unplanned Downtime

Along with improving resilience, AI and machine learning technologies are enabling us to improve the performance and competitiveness of GE's energy businesses and our customers. We're using AI and machine learning to improve GE's energy services portfolio, helping our customers optimize and reduce unplanned downtime with their power assets.

Through GE's Asset Performance Management platform, we can help our customers avoid disruptions by providing deep, real-time data insights on the condition and operations of the plant while at the same time factoring in predictive, forward looking forecasts for energy demand, weather and other factors that operators should account for to optimize overall plant management. This is another example that illustrates how critical it is for the digital solutions using AI and machine learning be linked to concrete physics-based models and data.

Using AI, machine learning, and digital twins, we can better predict when critical assets require repair or have a physical fault. This allows our customers to move towards a condition- based vs. schedule- based maintenance system, meaning assets can be brought in only when a repair is needed. For example, we can use machine learning to quickly and efficiently compare how a machine is operating versus known, standard operations. If there are deviations between the two, critical repairs can be identified, prioritized and compared to previously known system outage schedules allowing a more optimized maintenance planning process.

Recommendations

The examples I have shared and GE's extensive developments with AI and machine learning for complex industrial systems like power plants have given us firsthand experience and insights into what it takes to successfully integrate these technologies into our nation's critical energy infrastructure. With this perspective, I share the following recommendations:

1. ***Continue funding opportunities for public/private partnerships to expand the application and benefits of AI and machine learning across the energy system.*** The Digital Ghost/cybersecurity solution discussed earlier pertained only to power plants and just scratches the surface of neutralization. More research is needed to get to a mature cyber-attack resilient controls system for power plants. We think the application of this solution could be expanded to more distributed energy systems that involve renewable and gas power as well. We encourage the Congress to continue funding opportunities through the DOE for public/private partnerships with industry to bolster cybersecurity protections for critical power infrastructure.

2. ***Present and future R&D programs in AI and machine learning require both physical and digital experts.*** In the consumer internet, the value of data is measured in quantities not quality. For example, with online retail and advertising companies, it's the increasing quantity of data on people's shopping or search habits that supports their business models by allowing them to more accurately predict what any one individual might want to buy or purchase. In the industrial space, it is critical to match your digital solutions to the true physics of a physical machine or system. It's about finding the right data that helps you achieve a desired business outcome. As mentioned in the Digital Ghost/cybersecurity example, the ability to detect, localize and neutralize cyberthreats is tied to the physical understanding of the power plant.

3. ***Continue to invest in the nation's high- performance computing (HPC) assets and expand opportunities for private industry to work with the National Labs.*** The DOE and National Labs have done a tremendous job with maintaining and enhancing the nation's high-performance assets and in creating more collaborative opportunities with industry. The incredible computing power these systems offer is a powerful tool for industry in scientific discovery.

   We are still working to fully leverage our current computing capability, but we are certain that continued advances - particularly in exascale and quantum computing - will serve to create new advances in AI and machine learning – a potential source of competitive advantage for US industry. It also will help accelerate future industrial applications of AI and machine learning. Quite simply, it offers a competitive advantage for US industry.

Conclusion

We appreciate the opportunity to testify and offer our perspective on how the development of AI and machine learning technologies can meet the shared goals of creating a more efficient and resilient energy infrastructure.

One final thought is to reinforce a theme I have emphasized throughout my testimony, and that is the importance of having teams of physical and digital experts involved in driving future AI and machine learning solutions. I can personally speak to this firsthand, being a scientist that started my career as a physicist and is today applying my physical expertise in a digital role building digital twins of industrial assets.

In our opinion, you can have the best software developer in the world, but their solutions won't deliver the intended business outcome unless it's tied to real physical data and industrial domain knowledge to guide it.

Thank you and I look forward to answering any questions.

Matthew Nielsen
Principal Scientist
GE Global Research Center
Niskayuna, New York

Matt was born and raised in Michigan, where he attended undergraduate school at Alma College. In 1999, he received his PhD in Physics from Rensselaer Polytechnic Institute (RPI), located in Troy, NY.

During his dissertation research, Matt worked with General Electric to help develop electronic materials for high frequency RF applications. After graduation, he was able to join full-time and worked on a variety of efforts from electronic packaging to wide band gap semiconductors. Matt later led a large research program developing technology in the area of photonics, more specifically ultra-fast optical communications and three-dimensional optical storage materials and systems. Matt was also Lab Manager for the Electrophysics and Materials organization, where he initiated new efforts in bio-electronics and monitoring applications.

Matt currently has the position of Principal Scientist with a research focus on Digital Twin and specifically computer modeling/simulation of physical assets, using first-principle physics and machine learning methods. Applications using the Digital Twins include performance optimization and cyber-security.

1 Research Circle, K1 C38A
Niskayuna, New York 12309
T 518 387-4233
C 518 867-9202
E nielsema@ge.com

Chairman WEBER. Thank you, Dr. Nielsen.
Dr. Rollett, you're recognized for five minutes.

TESTIMONY OF DR. ANTHONY ROLLETT, U.S. STEEL PROFESSOR OF MATERIALS SCIENCE AND ENGINEERING, CARNEGIE MELLON UNIVERSITY

Dr. ROLLETT. So my thanks to Chairman Weber, Chairman Smith, Chairwoman Comstock, Ranking Members Veasey and Lipinski, and all the Members for your interest.

Speaking as a metallurgist, it's my pleasure and privilege to testify before you because I've found big data and machine learning, which depend on advanced computing, to be a never-ending source of insight for my research, be it on additive manufacturing or in developing new methods of research on structural materials.

My bottom line is that there are pervasive opportunities, as you've heard, to benefit from big data and machine learning. Nevertheless, there are many challenges to be addressed in terms of algorithm development, learning how to apply the methods to new areas, transforming data into information, upgrading curricula, and developing regulatory frameworks.

New and exciting manufacturing technologies such as 3–D printing are coming on stream that generate big data, but they need further development, especially for qualification, in other words, the science that underpins the processes and materials needed to satisfy requirements.

So consider that printing a part with a powder bed machine, standard machine, requires 1,000-fold repetition of spreading a hair's-breadth layer of powder, writing that desired shape in each layer, shifting the part by that same hair's breadth, and repeating. So if you think about taking a part and dividing the dimension of that part by a hair's breadth, multiplied by yards of laser-melting track, you can easily estimate that each part contains miles and miles of tracks, hence, the big data.

The recent successes with machine learning have used data that is already information-rich, as you've heard, cats, dogs, and so on. And so to advanced manufacturing and basic science, however, we have to find better ways to transform the data, stream into a big information stream.

Another very important context is that education in all STEM subjects needs to include the use of advanced computing for data analysis and machine learning. And I know that this Committee has focused on expanding computer science education, so thank you for that.

So for printing, please understand that the machines are highly functional and produce excellent results. Nevertheless, if we're going to be able to qualify these machines to produce reliable parts that can be used in, for example, commercial aviation, we've got some work to do.

If I might ask for the video, Daniel, if you can manage to get that to play. So I'd like to illustrate the challenges in my own research.

[Video shown.]

Dr. ROLLETT. I often used the light sources, in other words, x-rays from synchrotrons, most of which are curated by the Department of Energy. I use several modes of experimentation such as computer topography, diffraction microscopy, and dynamic x-ray radiography. So this DXR technique produces movies of the melting of the powder layers exactly as it occurs in 3–D printing with the laser. And again, at the micrometer scale you can see about a millimeter there. And you can also see that the dynamic nature of the process means that one must capture this at the same rate as, say, the more familiar case of a bullet going through armor.

Over the last couple of years, we've gotten many deep insights as to how the process works, but again, for the big-data aspect, each of these experiments lasts about a millisecond. That's about 500 times faster than you can blink. And it provides gigabytes of images, hence, the big data. Storing and transmitting such large amounts of data, which are arriving at ever-increasing rates, is a challenge for this vital public resource. I should say that the light sources themselves are well aware of this challenge. Giving more serious attention to such challenges requires funding agencies to adopt the right vision in terms of recognizing the need for fusion of data science with the specific applications.

I also want to say that cybersecurity is widely understood to be an important problem with almost weekly stories about data leaks and hacking efforts. What's not quite so well understood is exactly how we're going to interface manufacturing with cybersecurity.

So, in summary, I suggest that there are three areas of opportunity. First, federal agencies should continue to support the application of machine learning to advanced manufacturing, particularly for the qualification of new technologies and materials. I thank and commend all of my funders for supporting these advances and particularly want to call out the FAA for providing strong motivation here.

In the future, research initiatives should also seize the potential for moonshot efforts on objectives such as integrating artificial intelligence capabilities directly into advanced manufacturing machines and advancing synergy between technologies such as additive manufacturing and robotics.

Second, we need to continue to energize and revitalize STEM education at all levels to reflect the importance of the data in learning and computing with a focus on manufacturing. I myself have had to learn these things as I've gone along.

Third, based on the evidence that machine learning is being successfully applied in many areas, we should encourage agencies to seek programs in areas where it's not so obvious how to apply the new tools and to instantiate programs in communities where data, machine learning, and advanced computing are not yet prevalent.

Having traveled abroad extensively, I can assure you that the competition is serious. Countries that we used to dismiss out of hand, they're publishing more than we are and securing more patents than we do.

Again, I thank you for the opportunity to testify and share my views on this vital subject. I know that we will all be glad to answer your questions.

[The prepared statement of Dr. Rollett follows:]

Testimony

Before the Committee on Science, Space, and Technology, Subcommittee on Research and Technology and Subcommittee on Energy, of the U.S. House of Representatives on the hearing titled, "Big Data Challenges and Advanced Computing Solutions"

Submitted By

Anthony David Rollett

US Steel Professor of Materials Science and Engineering

Carnegie Mellon University

July 12th, 2018

Greetings, my name is Anthony Rollett and I am a Professor of Materials Science & Engineering in the College of Engineering at Carnegie Mellon University, in Pittsburgh, PA. At Carnegie Mellon I help lead the NextManufacturing Center, which is focused on advancing additive manufacturing and participate in the Manufacturing Futures Initiative—a campus wide effort focused on accelerating innovation and enhancing manufacturing in the Greater Pittsburgh region.

I thank Chairman Weber of the Energy Subcommittee and Chairwoman Comstock of the Research and Technology Subcommittee for inviting me to testify today. I also thank, Ranking Member Veasey, Ranking Member Lipinski and all of the Members of the Committee for your interest in Big Data and Computing, which are subjects of great interest and importance in my research.

In recent years, I have had the privilege of becoming closely involved in research on 3D printing, which is a key component of advanced manufacturing. It is clear to me that this is a seriously revolutionary technology because it forces us to think differently about how to make things. The design of a part is as intimately coupled to the printing process and the chosen material as a Stradivarius is to its wood and crafting. The difference is the importance of data as input and as output. Imagine that in a few years we will be able to, e.g., build a rocket that is tailored to the particular mission, instead of forcing the payload to match one of a limited set of vehicles. Or that "mass production" is transmuted into "mass individualization" such that Ford's proverbial "any color so long as black" becomes "any choice of color and size for dozens if not hundreds of parts of a car."

Let me begin by giving some context to the challenges and solutions by explaining that there are both practical applications and scientific advances to be gained from adapting to the availability of large amounts of data. I will outline three major challenges in advanced manufacturing, associated needs in STEM education, and a link between cybersecurity and manufacturing.

New and complex manufacturing technologies such as 3D printing are coming on stream that need further development, especially for qualification—the science of verifying that processes and/or materials will produce the characteristics required. Scientific research is generating ever larger streams of data that are challenging to handle and to interpret. There is an essential challenge in this that concerns the extraction of information from data. Current machine learning algorithms have been developed for big data, as we know, but the data concerned is information rich, e.g., faces or cats or cars. One could say that "big information" is as important as "big data."

Another important context is that education in all STEM subjects needs to include the use of advanced computing for data analysis. Our data acquisition instruments are essentially all run by computers and modeling our experiments essentially always involves large computer codes. These considerations and the examples that follow demonstrate the pervasive nature of big data and computing and the need to incorporate computer-aided learning into every aspect of how we function in science and engineering. I know that this Committee has focused on expanding computer science education. Those efforts are critical and appreciated.

Returning to the domain of advanced manufacturing, another exciting area of the application of big data, machine learning and advanced computing is to the manifold challenges of the materials used. In order to check the internal structure of materials and to understand their properties, we commonly cut, polish and photograph cross-sections.

An early lesson for those of us in materials science and engineering was to be told that although computer vision is well developed for finding cats, dogs, cars etc. in images or videos, the sort of cross-sectional images that we produce are for more complex and the features in each image much less obvious. An analogy might be a painting by Jackson Pollock with its seemingly analysis-defying random dribbles of paint. Yet there is an organization to the image that makes his paintings compelling art. I assert that many of the images of materials are equally complex and require domain-specific analysis. We have demonstrated success with examples of teaching the computer to recognize different kinds of metal powders used in additive manufacturing and recognizing different kinds of steel where the composition is constant but the processing history is varied. The bottom line is that advanced manufacturing is already a source of big data but there are many challenges in front of us to learn how to transmute the data into information and then discover the algorithms that allow us to learn from that data and optimize manufacturing.

These successes open up a wide range of potential impacts on improving materials, generating new materials, performing quality control on feedstocks, etc. We have also demonstrated that we can recognize failures in any individual powder spreading step that is essential to any powder bed 3D printing process; again, this points to an impact of improved machine control algorithms that exploit data, machine learning and high speed computing.

As another example of how data coupled with machine learning shows up almost everywhere, I ran across the "LettuceBot," which is a software that controls an agricultural machine towed by a tractor across lettuce plantations whose job is to decide which individual lettuce plants should stay in the ground to grow versus those that should be culled. Once again, cameras provide images that are then analyzed for a decision-making process followed by action (or not). The bigger picture is one where computer vision helps humans to make decisions about the manufacturing process that they are in charge of, i.e., advanced technology aiding workers .

To illustrate the challenges in my own research, I often use the light sources, i.e., x-rays from synchrotrons, most of which are curated by the Department of Energy. I use several modes of experiment such computed tomography, high energy diffraction microscopy and dynamic x-ray radiography. Computed tomography (CT) is very similar to getting an MRI scan except that the high energy x-rays from the synchrotron allows one to see inside the sort of dense materials from which we build aircraft, engines, rockets etc. at the micrometer scale (about 1/100th of a human hair). The results have been invaluable in understanding the feedstocks used in, e.g., 3D metals printing. High energy diffraction microscopy (HEDM) functions as a microscope, again at the micrometer scale, that provides, in full 3-dimensional form, a map with crystal information of the material in millimeter-sized samples. We can then heat and deform the sample and measure how it responds under load, which is proving invaluable for understanding what controls the durability of components, for example. Once again, a central challenge is how to transmute this ever expanding data stream into information and to discover the algorithms that allow us to learn from that data.

Dynamic x-ray radiography (DXR) provides movies of the melting of powder layers just as occurs in 3D printing with a laser, again at the micrometer scale. The dynamic nature of the process means that one must capture the process at the same rate as the more familiar case of a bullet penetrating armor. Over the last couple of years this technique has provided many deep insights into how additive manufacturing really works at the appropriate scale of length and time. From the perspective of data and computing, each experiment lasts for about a millisecond, i.e., 500 times faster than you can blink, and provides gigabytes of images. It is not difficult , therefore, to appreciate that a few days' worth of what we call "beam time" provide big data, so much so that we typically take it home on our own hard disk drives. Storing such large amounts of data is a challenge for this extremely important public resource. Transmitting such large amounts of data, e.g., to one's own university, is challenging. The mechanisms exist but they are not quick as they ought to be and accessing high speed transfers is definitely something for experts, even if the institutions at either end have the appropriate speed of access. The light sources themselves are well aware of the forthcoming challenge that is posed by the rapidly accelerating rate at which data is generated in the aggregate as detectors become ever larger and more sensitive.

Acquiring the data may only occupy a few days, but analyzing it often consumes months of time on the part of a graduate student. Perhaps a better way to say this is that we have the problem of converting raw data to useful information, by contrast with the data available from, e.g., social media, which are intrinsically information-rich. My judgment is that advanced computing algorithms to aid researchers in the conversion of data to information are nascent at best. Although there is a plethora of data analytics and machine learning techniques available, applying such techniques in any given domain requires time and effort.

Giving more serious attention to such challenges requires funding agencies to adopt the right vision in terms of recognizing the need for a fusion of research activities. We are in essence building the infrastructure for digital engineering and manufacturing.

A closely related issue is the timescale on which new methods are developed. The canonical 3 or 4-year research program rarely allows one to take a technique development to a reasonable point of maturity or technical readiness level in the modern argot. The high energy diffraction microscopy mentioned above is a case in point where an agency sustained the effort over roughly a decade, which enabled it to mature to the point where the research community was able to start using it more generally.

Additive manufacturing provides an excellent example of an application domain for big data and computing. Consider 3D printing of metals as a particular facet that has grown with dramatic speed from a small specialized activity that most believed (as did we) would only provide business cases in aerospace and only in rare instances, to a technology that essentially all OEMs consider that they must pay attention to. It is also provoking a reaction in education, where universities are acting at something faster than the proverbial glacial pace and instituting new programs across the scale, e.g., MS programs in additive manufacturing.

To print a part with a powder bed machine requires thousands-fold repetition of spreading a hairsbreadth layer of powder, writing the desired shape in that layer, shifting the part by the hairsbreadth, and repeating. Divide a part dimension by a hairsbreadth, multiply by yards of laser melt track, and one readily estimates that each part contains miles upon miles of melt tracks. There is a great deal of physics and chemistry detail required at the melt track scale.

Thus, the data stream is commensurately enormous ("big"), but the impact has to be such that useful information about the integrity of the part is obtained. Please do not be intimidated by the scale because the machines are highly functional and produce good results. Nevertheless, if we are to be able to qualify the machines to produce reliable parts that can be used in, e.g., commercial aviation, there is work to do.

Moreover if we as a country are to maintain our competitiveness in this area, we need the full range of tools that, crucially, include the application of big data and advanced computing. As a brief illustration, consider taking high speed movies of the melt pool using visible light (as opposed to the highly specialized x-ray approach). This generally has to be done at an angle to the laser beam and the images are confused by particle spatter and metal vapor plumes. This means that substantial processing must be done on the videos to render them useful to the researcher. We are only at the very beginning of being to use this type of data, let alone knowing how to incorporate the lessons learned as improved control algorithms. Permit me to underscore the importance of the research community and the publication of results so that companies involved in advanced manufacturing can adopt the results without necessarily revealing where they obtained the knowledge.

Finally, cybersecurity is widely understood to be an important problem, with almost weekly stories about data leaks and hacking efforts. What is less well understood is how manufacturing and cybersecurity must interface to each other. At the consumer level, concern has already been expressed about the ability of bad actors to gain access to IoT-enabled gadgets in one's home and control them or acquire data from them. With companies touting their ability to provide customer solutions that are based on networked machines, the importance of cybersecurity in manufacturing takes on a new significance and urgency. The caution in this instance is to not underestimate the importance of the domain-specific knowledge for determining which existing cybersecurity solutions will work and, more importantly, adapting the methods to suit a given domain. This is analogous to the way in which computer vision is applicable to materials science but has to be adapted to the particularities of the field.

Recommendations

As others have testified, the various agencies that provide federal funding for R&D have done an excellent job over the years of identifying worthwhile areas for development of new ideas. Please continue to support them.

Specialized facilities are tremendously important to the scientific and engineering community. The DOE has done an exceptional job in this regard and my own research is all the richer for it. In addition, DOE is investing in building machine learning capabilities. The manufacturing institutes—such as America Makes—have also been critical to advancing more applied research.

I suggest that there are three areas of opportunity. First, federal agencies should continue to support the application of machine learning to advanced manufacturing particularly for the qualification of new technologies and materials. Currently, no additive manufacturing processes or materials are qualified for mission critical defense or aerospace parts (non-mission critical additive parts are in use). As noted above, this requires advances in scientific research and strong collaboration with industry and among research and application and regulatory agencies. Winning the innovation race in the science of qualification is essential for future competitiveness and job creation in these technologies. In the future, research initiatives can also seize the potential for "moonshot" efforts on objectives such as integrating artificial intelligence capabilities directly into advanced manufacturing machines and advancing synergy between technologies such as additive manufacturing and robotics.

Second, we need to continue to energize and revitalize STEM education at all levels to reflect the importance of data, learning and computing, with a focus on manufacturing. Data analytics will play a vital role across the entire manufacturing enterprise—from the lab, to product design, production and product service functions. It will not be necessary for all workers to have a computer science degree. But varying degrees of comfort and capability with statistics and data analytics will be vital. As a step in this direction, the NextManufacturing Center at Carnegie Mellon has begun engaging teachers and students with the most advanced additive manufacturing machines. Investments that creatively stimulate a co-development of manufacturing with cybersecurity innovation will be essential.

Third, based on the evidence that machine learning is being successfully applied in many areas, we should encourage agencies to seek programs in areas where it is not so obvious how to apply the new tools and to to instantiate programs in communities where data, machine learning and advanced computing are not yet prevalent. Not only is domain-specific knowledge essential but, in manufacturing and research, the process of transmuting data into knowledge is a challenge in itself. In fact, one could say that "big information" is the twin of "big data."

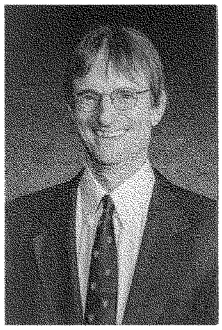
Having traveled abroad extensively, I can report that the competition in science and technology is serious. Countries that we used to dismiss out of hand are publishing more than we are and securing more patents than we do. Time and again, national investment in new ideas and technology, coupled with an expectation that industry will strive to pick these up in their innovation process, has kept this country in the lead.

A best practice in my experience is where a funding agency has a well-established mechanism through which the scientific community can articulate needs and directions. Although a variety of mechanisms is appropriate, some work better than others. It is important that the community recognizes and is comfortable with whatever mechanism an agency uses.

Program Managers should have some discretion in what they fund so that they are able to respond quickly when an interesting new idea arises. High risk with high impact is often touted but less often encouraged.

Although some effort has been made to facilitate the transport of big data around the country, arranging to ship data at the terabyte scale requires substantial effort for ordinary researchers. This is, of course, is linked to the availability of data storage systems ("servers") that have the capacity and delivery speed to support such transfers. It would be helpful if one of the agencies were to be empowered to support such capabilities.

Again, thank you very much for the opportunity to share my views on this vital subject. I would be glad to answer any follow up questions you may have.



Rollett's research focuses on microstructural evolution and microstructure-property relationships in 3D, using both experiments and simulations. Interests include 3D printing of metals, materials for energy conversion systems, strength of materials, constitutive relations, microstructure, texture, anisotropy, grain growth, recrystallization, formability, extreme value statistics and stereology. Relevant techniques highlight spectral methods in micro-mechanics, Dynamic X-ray Radiography and High Energy Diffraction Microscopy. Important recent results include definition of process windows in 3D printing through characterization of porosity, 3D comparisons of experiment and simulation for plastic deformation in metals, the appearance of new grains during grain growth, and grain size stabilization. He has been a Professor of Materials Science & Engineering at Carnegie Mellon University since 1995 and before that was with the Los Alamos National Laboratory. His most recent honor was the award of US Steel Professor of Metallurgical Engineering & Materials Science in 2017. He is the co-Director of CMU's NextManufacturing Center that is dedicated to advancing manufacturing especially through 3D printing. He has over 200 peer-reviewed publications

Chairman WEBER. Thank you, Doctor. I now recognize myself for five minutes.

This question is for all the witnesses. You've all used similar terminology in your testimonies like artificial intelligence, machine learning, and deep learning. So that we can all start off on the same page, I'll start with Dr. Kasthuri. But could you explain what these terms mean and how they relate to each other?

In the interest of time, I'm going to divvy these up. Dr. Kasthuri, you take artificial intelligence. Dr. Yelick, you take machine learning. Dr. Nielsen, you take deep learning. All right? Doctor, you're up.

Dr. KASTHURI. Thank you, Chairman Weber. That's an excellent question. In the interest of time I'm not going to speak about artificial intelligence. There are clearly experts sitting next to me. I'm interested in the idea of finding natural intelligence wherever we can, and I would say that the confusion that exists in these terminologies also exist when we think about intelligence beyond the artificial space. And I'm happy to—maybe perhaps after I let the other scientists speak to talk about how we define natural intelligence different ways, which might help elucidate the ways we define artificial intelligence.

Chairman WEBER. All right. Fair enough. Dr. Yelick, do you feel that monkey on your back?

Dr. YELICK. Yes. Thank you very much for the question. So let me try to cover a little bit of all three. So artificial intelligence is a very long-standing subfield of computer science looking at how to make computers behave with humanlike behavior. And one of the most powerful techniques for some of the subproblems in artificial intelligence such as computer vision and speech processing are machine-learning algorithms. These algorithms have been around for a long time, but the availability of large amounts of labeled data and large amounts of computing have really made them take off in terms of being able to solve those artificial intelligence problems in certain ways.

The specific type of machine learning is a broad class of algorithms that come from statistics and computer science, but the specific classes called deep learning algorithms, and I won't go into the details. I will defer that if somebody else wants to try to explain deep learning algorithms, but they are used for these particular breakthroughs in artificial intelligence.

I would say that the popular press often equates the word artificial intelligence with the term deep learning because the algorithms have been so powerful, and so that can create some confusion.

Chairman WEBER. All right. Thank you. Dr. Nielsen?

Dr. NIELSEN. Yes, I'm not an expert in deep learning, but we are practitioners of deep learning at GE. And really it's taken off in, I would say, the last several years as we've seen a rise in big data. So we have nearly 300,000 assets spread globally and each one generating gigabytes of data. Now, processing that gigabytes of data and trying to make sense of it we're using deep learning techniques. It's a subfield, as you mentioned, of machine-learning algorithms but allows us to extract more information, more relationships if you will.

So, for example, we use deep learning to help us build a computer model of a combined-cycle power plant, very complex system, very complex thermodynamics. And it's only because we have been able to collect now years and years of historical data and then process it through a deep-learning algorithm. So, for us, deep learning is a breakthrough enabled by advances in computing technology, advances in big-data science, and it's allowing us to build what we think is more complex models of not only our assets but the processes that they perform.

Chairman WEBER. And, Dr. Rollett, before you answer, you issued a warning quite frankly in your statement that there's been more patents filed by some of the foreign countries than we are. Do you attribute that to what we're talking about here? Go ahead.

Dr. ROLLETT. In very simple terms, I think what I'm calling attention to is investment level in the science that underpins all kinds of things, so whether it be the biology of the brain, the functioning of the brain or how you make machines work, how you construct machines, control algorithms, so on, and so forth. That's really what I'm trying to get at.

Chairman WEBER. Okay.

Dr. ROLLETT. And I'm trying to give you some support, some ammunition that what you're doing as a committee, set of Subcommittees is really worthwhile.

Chairman WEBER. Yes, well, thank you. I appreciate that.

I'm going to move on to the second question. Several of you mentioned your reliance on DOE facilities, which is, again, what you're talking about, particularly light sources and supercomputing which we are focused on, have been to a couple of those for the types of big-data research that you all perform and my question is how necessary is it for the United States to keep up to date? You've already address that with the patents statement, a warning that you issued, but what I want to know is have any of you all—would you opine on who the nearest competitor is? And have you interfaced with any scientists or individuals from those companies? And if so, in what field and in what way? Doctor?

Dr. KASTHURI. I would say that, internationally, sort of the nearest two competitors to us are Germany and China. And in general in the scientific world there is a tension between collaboration and competition independent of whether the scientist lives in America or doesn't live in America.

I think the good news is that for us at least in neuroscience we realize that the scale of the problem is so enormous and has so much opportunity, there's plenty of food for everyone to eat. So right now, we live at the world of cooperation between individual scientists where we share data, share problems, and share solutions back and forth unless of course familiar with what happens at levels much higher than that.

Chairman WEBER. Thank you. Dr. Yelick?

Dr. YELICK. Yes, in the area of high-performance computing I would say the closest competitor at this point is China. And in science we also like to look at derivatives, so what we really see is that China is growing very, very rapidly in terms of their leadership. At this point we do have the fastest computer and the top-500 list in the United States, but of course until recently that was the top two—the number-one and -three machines were from China. But perhaps more importantly than that there are actually more machines manufactured in China on that list than there are machines that are fractured in the United States, so there is a huge and growing interest, and certainly a lot of research, a lot of funding in China for artificial intelligence, machine learning, and all of that applied to science and other problems.

Chairman WEBER. Have you met with anybody from over in China involved in the field?

Dr. YELICK. Yes. Last summer, I actually did a tour of all of the major supercomputing facilities in China, so I got to see what were the number-one and number-three machines at that time—and was very impressed by the scientists. I think one of the things that you see—and a lot of, by the way, very junior scientists, the students that they are training in these areas, they use these machines to also draw talent back to China from the United States or to keep talent that was trained in China in the United States. And they have very impressive people in terms of the computer scientists and computational scientists.

Chairman WEBER. And, Dr. Nielsen, very quickly because I'm out of time.

Dr. NIELSEN. Yes, I would just like to echo that, like Dr. Rollett, we follow publications and patents, and we're seeing a growing number from China, so I'd like to echo that just from that statement. We're seeing growing interest in the use of high-performance computing to go look at things like cybersecurity from China, so obviously, that's the number-one location we're looking at.

Chairman WEBER. Good. Thank you, Dr. Rollett. I'm happy to move on now. So I'm now going to recognize the gentlelady from Oregon for five minutes.

Ms. BONAMICI. Thank you very much, Mr. Chairman.

What an impressive panel and what a great conversation and an important one.

I represent northwest Oregon where Intel is developing the foundation for the first exascale machines. We know the potential of high-performance computing and all energy exploration, predicting climate weather, predictive and preventive medicine, emergency response, just a tremendous amount of potential. And we certainly recognize on this Committee that investment in exascale systems and high-performance computing is important for our economic competitiveness, national security, and many reasons.

And we know—I also serve on the Education Committee, and I know that our country has some of the best scientists and programmers and engineers, but what really sets our country apart is entrepreneurs and innovation. And those characteristics require creative and critical thinking, which is fostered through a well-rounded education, including the arts.

I don't think anyone on this Committee is going to be surprised to hear me mention the STEAM Caucus, which is—I'm cochairing with Representative Stefanik from New York, working on integrating arts and design into STEM, learning to educate innovators. We have out in Oregon this wonderful organization called Northwest Noggin, which is a collaboration of our medical school, Oregon Health Sciences University, Portland State University, Pacific

Northwest College of Art, and the Regional Arts and Culture Council. And they go around exciting the public about ongoing taxpayer-supported neuroscience research. And they're doing great work and expanding the number of people who are interested in science and also communicating with all generations and all people about the benefits of science.

So, Dr. Rollett, in your testimony you talked about the role of data analytics across manufacturing—the manufacturing sector. And you noted that it's not necessarily going to be important for all data analytic workers to have a computer science degree, so what skills are most important for addressing the opportunities? You did say in your testimony that technology forces us to think differently about how to make things, so talk about the next manufacturing center at Carnegie Mellon and what you're doing to prepare students for evolving fields? And we know as technology changes we need intellectual flexibility as well, so how do you educate people for that kind of work?

Dr. ROLLETT. So thank you for the opportunity to address that. The way that we're approaching that is telling our students don't be afraid of these new techniques. Jump in, try them, and lo and behold, almost every time they're trying it—sometimes it's a struggle, but almost every time that they try it they're discovering, oh, this actually works. Even if it's not big data in quite the sense that, say, Kathy would tell us, even small data works.

So, for example, in these powder bed machines you spread a layer. Well, if you just take a picture of that layer and then another picture and you keep analyzing it and you use these computer vision techniques, which are sort of a subset of machine learning, lo and behold, you can figure out whether your part is building properly or not. That's the kind of thing that we've got to transmit to all of our students to say it's not that bad, jump in and try it and little by little, you'll get there.

Ms. BONAMICI. I think over the years many students have been very risk-averse and they don't want to risk taking something where they might not get the best grade possible, so we have to work on overcoming that because there's so much potential out there until students have the opportunity to get in and have some of that hands-on learning.

Dr. Yelick, I'm in the Northwest and it's not a question of if but when we have an earthquake off the Northwest coast, and a tsunami could be triggered of course by that earthquake along the Cascadia subduction zone. So in your testimony you discuss the research at Berkeley Lab to simulate a large magnitude earthquake, and I listened very carefully because you were talking about the effects on an identical building in different areas. This data could be really crucial as we are assessing the need for more resilient infrastructure not only in Oregon but across the country. So what technical challenges are you facing and sort of curating, sharing, and labeling and searching that data? And what support can the federal government provide to accelerate a resolution of these issues?

Dr. YELICK. Well, thank you very much for the question. Yes, this is very exciting work that's going on, and simulating earthquakes is currently at a regional scale. There are technology challenges to trying to even get that to larger-scale simulations, but I think even more importantly the work that I talked about is trying to use information about the geology to try to give you much more precise information about the safety of a particular location.

And the challenge is to try to collect this data and then to actually invert it, that is turn it into a model so you collect the data and then in some sense you're trying to develop a set of equations that say how that area—based on the data that's been collected from little tiny seismic events, it'll tell you something about how that particular subregion, even a yard or a city block or something like that, how that city block is going to behave in an earthquake. And you can use the information from tiny seismic events and then to infer how it will behave in a large significant earthquake. And so there's technical challenge, mathematical challenges of doing that, as well as the scale of computing for both doing the data, inverting the data but also then doing the simulation.

And I think you bring up a very good point about the community needs for these community data sets because you really want to make it possible for many groups of people, not just, for example, a power company that has smart meter data but for other people to access that kind of data.

Ms. BONAMICI. Thank you. And I want to follow up with that. I'm running out of time, but as we talk about infrastructure and investment in infrastructure, we know that by making better decisions at the outset we can save lives and save property, so the more information we have about where we're building and how we're building is going to be a benefit to people across this country, as well as in northwest Oregon. So thank you again to this distinguished panel. I yield back.

Chairman WEBER. Thank you, ma'am.

The gentlelady from Virginia, Mrs. Comstock, is recognized.

Mrs. COMSTOCK. Thank you, Mr. Chairman, and thank all of you here. This has been very interesting once again.

Now, I guess I'd ask to all of you, what are the unexamined big-data challenges that could benefit from machine learning? And what are the consequences for the United States for not being the world leader in that if we aren't going forward in the future? Maybe, Dr. Rollett, if you'd like to start. You look like you had an answer ready to go, so——

Dr. ROLLETT. I'll give you a small example from my own field. So when we deal with materials, then we have to look inside the materials. So we typically take a piece of steel and we cut it and we polish it and we take pictures of it. So traditionally, what we've done is play the expert witness as it were. You look at these pictures, which I often say resemble more of a Jackson Pollock painting than anything that remotely as a simple as a cat, and so the excitement in our field is that we now have the tools that we can start to tease things out of these pictures, that we go from something where we are completely dependent on sort of gray-bearded experts to let the computer do a lot of the job for you. And that speeds things up and it automates them and it allows companies to detect problems that they're running across. So it's just one example.

Dr. KASTHURI. Congresswoman Comstock, thank you for the question. I have two sort of answers specifically to thinking about brains and then to thinking about education. I think these are the potential things that we can lose. One of the things that I find fascinating about how our brains work is that whether you are Einstein thinking up relativity or Mozart making a concerto or you're just at home watching reality TV, all brains operate at about 20 watts of energy. These light bulbs in this room are probably at 60 watts of energy. And although you might already think some of your colleagues are dim bulbs, in this sense, what's amazing about the things that they can accomplishes that they accomplish them at energy efficiencies that are currently unheard of for any type of algorithm.

So I feel like if we can leverage machine learning, deep analytics, and understand how the brain passes information and processes information for energies that are really energy efficiencies unheard of in our current algorithms and robots, that's a huge benefit to both the national and economic securities of our country. That's the first.

And the second thing I'd like to add, the other reason that it's important for us to lead now—and I'll do it by example—is that in 1962 at Rice University John F. Kennedy announced that we were going to the moon. And he announced it and in his speech he said we're going to go to the moon—and I paraphrase—not because it's easy but because it's hard and because hard things test our mettle and test our capabilities.

The other interesting fact about that is that in 1969 when we landed on the moon, the average age of a NASA scientist was 29 years old, so quick math suggests that when Kennedy announced the moonshot, many of these people were in college. They were students. And there was something inspirational about positing something difficult, positing something visionary. And I suspect that this has benefited us—in recruiting this generation of scientists to the moonshot has benefited this country in ways that we yet haven't calculated. And I suspect that if we don't move now, we lose both of these opportunities, among many others.

Mrs. COMSTOCK. So it's really a matter of getting that focus and attention and commitment so that you have that next generation understanding this is really a long-term investment, and we have a passion for it, so they will.

Dr. KASTHURI. Exactly.

Dr. YELICK. I'll just add briefly that I think we really want to—in terms of the threat associated with this is really about continuing to be a leader in computing but also about the control and use of information. And you can see the kinds of examples we've given are really important, and you hear about it in the news about the control and use of information. We need leaders in understanding how to do that and make sure that information is used wisely.

We teach our freshmen at Berkeley a course in data science, so whether they're going to go off and become English majors or art majors or engineers, we think it's really important for people to understand data.

Dr. NIELSEN. And just real briefly, I'd like to build a little bit on Dr. Rollett's comments. For us, we're seeing tremendous benefit in big data for things like trying to better predict when an aircraft engine part has to be repaired, when it needs to be inspected, very critical for the safety of that engine. For gas turbines, same thing. Wind parts need to be inspected and repaired.

So where does big data come in? It comes in with computational fluid dynamics, which we leverage—actually, the high-performance computing infrastructure of the United States materials science, material knowledge, trying to understand grain structure, et cetera. So for us, that nexus of the digital technologies with the physics, understanding the thermodynamics of our assets are leading us into what I think is just a better place to be from maintenance scheduling, safety, resiliency, et cetera.

Mrs. COMSTOCK. Thank you. I really appreciate all of your answers.

I yield back, Mr. Chairman.

Chairman WEBER. The gentleman from Virginia, Mr. Beyer, is recognized for five minutes.

Mr. BEYER. Mr. Chairman, thank you very much, and thank you all very much for doing this.

Dr. Kasthuri, so on the BRAIN Initiative I think obviously the most—maybe the most exciting thing happening in the world today, I was fascinated by this whole notion of the Connectome, 1 billion neurons with 1 quadrillion connections, you talk about it being if you took—of all the written material in the world into one data set, it'd just be a small fraction of the size of this brain map. Is it possible that it's simpler than that, that it sort of strains my understanding that there are few things in nature that are as complex as that. Why in evolution have we developed something that—and every human being on the planet has a brain that's already—contains more connections than every bit of written material?

Dr. KASTHURI. Congressman Beyer, that's a great question, and like most scientists I'm going to do a little bit of handwaving and a little bit of conjecture because the question that you're asking is the question that we are trying to accomplish. We know reasonably well that there are, as you said, 100 billion brain cells, neurons, that make on order 1 quadrillion connections in the brain. Now, that—when I say the data of that, I'm really talking about the raw image data. What will it take to take a picture of every part of the brain and if you added up all the data of all those pictures together, it would be the largest data set ever collected.

Now, I suspect we have to do that at least once and then it might be possible that there are patterns within that data that then simplify the next time that we have to map your brain. One way to think about this is that before we had a map of DNA, we didn't realize that there was a pattern within DNA, meaning every three nucleotides—A, C, T, et cetera—codes for a protein. And that essentially simplifies the data structure to, let's say, 1/3. I don't need to know, I just need to know that these three things are an internal pattern that then gets repeated again and again and again. And that was a fundamental insight. We have no similar insight into the brain. Is there a repetitive pattern that would actually reduce the amount of data that we had to collect?

So, you're right, it might be that the second brain or the third brain isn't going to be that much data, but now let me give you the counter because as a scientist I have to do both sides or all sides.

The other thing we know is that each human brain is unique, very much like a snowflake. Your brain, the connectivity, the connections in your brain at some level have to represent your life history, what your brain has experienced.

And so the question for me—and I think it's really one of the most important questions—is even within the snowflake there are things that are unique to snowflakes but they're the same. They either have seven arms are eight arms or six arms. I get them confused with spiders, but it's one of those is the answer. So there's regularity in a snowflake at the level of the arms, but there is uniqueness at the level of the things that jut out of the seven arms of the snowflake. And the fundamental question is what is unique, what is the part that makes each of us a neurological snowflake and what is common between all of us? And that would be one of the very first goals of doing a map is to discover the answer to your question.

Mr. BEYER. Yes, well, thank you for a very thoughtful answer. And I keep coming back to the Einstein notion that always looking for the simplest answers, things that unify it altogether. So here's another simple question. You talked in your very first paragraph about reverse engineering human cognition into our computers, good idea? At our most recent AI hearing here a lot of the controversy was, you know, dealing with Elon Musk and others and their concerns about what happens when consciousness emerges in machines.

Dr. KASTHURI. Again, a fantastic question. Here's my version of an answer. We deal with smarter things every day. Many of our children, especially mine, wind up getting consciousness and being smarter than us, certainly smarter than me, but yet we don't worry about the fact that this next generation of children, forever the next generation of children will always be smarter than us because we've developed ways as a society to instill in them the value systems that we have. And there are multiple avenues for how we can instill in our children the value systems that we have.

I suspect we might use the same things when we make smart algorithms, the same way we make smart children. We won't just produce smart algorithms but we'll instill in them the values that we have the same way that we instill our values in our children.

Now, that didn't answer your question of whether reverse engineering the brain is a specific good idea for AI or not. The only thing I would say is that no matter what we can imagine AI—artificial intelligence doing, there is a biological system that does that at more energy efficiency and its speed for which that AI physical silicon system does not. But I suspect these answers are probably best debated amongst you and then you could tell us.

Mr. BEYER. Well, that was a very optimistic thing. I want to say one of the things we do is we keep the car keys in those circumstances.

Mr. Chairman, I yield back.

Chairman WEBER. Thank you. The gentleman from Kansas is recognized for five minutes.

Mr. MARSHALL. Well, thank you, Mr. Chairman.

Speaking of Kansas, I'm sure you all remember President Eisenhower is the one who started NASA in 1958, but it was President

Kennedy, as several of you have stated, that, you know, gave us the definitive goal to get to the moon. And as a young boy I saw that before my eyes, the whole country wrapped around that.

Each of you get one minute. What's your big, hairy, audacious goal, your idea, it took 11 years, '58 to '69 to get to the Moon. Where are we going to be in 11 years? Dr. Rollett, we'll start with you and you each get one minute.

Dr. ROLLETT. I think we're going to see that manufacturing is a much more clever operation. It understands the materials. It understands how things are going to last, and it draws in a much wider set of disciplines than it currently does. I have to admit I don't exactly have an analogy to going to the moon, but that's a very good challenge.

Mr. MARSHALL. What I like about your idea is that's going to add to the GDP. Our GDP grows when we become more efficient, not when federal government sends dollars to States for social projects, so I love adding to GDP.

Dr. Nielsen, I guess you're next.

Dr. NIELSEN. So I would love it if every one of our assets—and I mentioned there are about 300,000 globally—had their own digital twin, so every aircraft engine had its own digital twin. A digital twin is a computer model that when the asset is operating, we're collecting data. So imagine an aircraft engine taking off. As soon as that aircraft engine takes off, we pull the data back from the aircraft engine and we update the computer model. That computer model becomes a digital twin of the physical asset. If every one of our 300,000-plus assets had a digital twin, we'd be able to know with very good precision when it needed to be maintained, when it needed to be pulled off wing, what kind of repairs when it went to a repair shop, what kind of repairs need to occur.

Mr. MARSHALL. You can do that with satellites and a whole bunch of things.

Dr. NIELSEN. We can pull back data from a whole variety of different pathways. It's then utilizing that data in the most efficient way, which we use machine learning and AI-type technologies——

Mr. MARSHALL. Maybe get internet to rural places by doing that, right?

Dr. NIELSEN. Yes.

Mr. MARSHALL. Okay. We better go on. Dr. Yelick?

Dr. YELICK. So I think one of the biggest challenges is understanding the microbiome and being able to use that information about the microbiome in both health applications and agriculture, in engineering, materials, and other areas.

So I think that we already know that your microbiome, your own personal microbiome is associated with things like obesity, diabetes, cardiovascular disease, and many other disorders. We don't understand it as well in agriculture, but we're looking at things like taking images of fields, putting biosensors into the fields and putting all this information together to understand how to make—to improve the microbiome to improve crop yield and reduce other problems. So I think it's about both understanding and controlling the microbiome, which is a huge computational problem.

Mr. MARSHALL. Okay. Dr. Kasthuri?

Dr. KASTHURI. The thing I would really like to have done in 11 years is understand how brains learn. And actually it reminds me of something that I should've said earlier about the differences between artificial intelligence, machine learning, deep learning, and how brains learn. The main difference is that for many of these algorithms you have to provide them thousands of examples, millions of examples, billions of examples before they can then produce inferences or predictions that are based on those examples.

For those of you with children, you know that that's not the way children learn. They can learn in one example. They can learn in half an example. Sometimes I don't even know where they're learning these things. And when they learn something, they learn not only the very specific details of that thing, they can immediately abstract it to a bunch of other examples.

For me, this happened with my son the first time he learned what a tiger was. An image of a tiger he could see, and then as soon as he learned that, he could see a cartoon of a tiger, he could see a tiger upside down, he could see the back of a tiger or the side of a tiger, and from the first example be able to infer, learn all of these other general applications.

If in 11 years we could understand how the brain does that and then reverse engineer that into our algorithms and our computers and robots, I suspect that will influence our GDP in ways that we hadn't yet imagined.

Mr. MARSHALL. Okay. Thank you so much. I yield back.

Chairman WEBER. I thank the gentleman.

The gentleman from the great State of Texas is recognized.

Mr. VEASEY. Thank you, Mr. Chairman.

Dr. Rollett, am I pronouncing that right?

Dr. ROLLETT. It'll do.

Mr. VEASEY. Okay. In your testimony you talk about the huge amounts of data that are generated by experiments using light sources to examine the processes involved in additive manufacturing. You also highlight the need for more advanced computing algorithms to help researchers extract information from this data. And you state that we are essentially building the infrastructure for digital engineering and manufacturing. I was hoping that you'd be able to expand on that a little bit and tell us also what are the necessary components of such infrastructure.

Dr. ROLLETT. Right. So one of the things that I didn't have time to talk about is where does the data go? And so, you know, one's generating terabytes, the standard story is you go to a light source, you do an experiment, all of that data has to go on disk drives, and then you literally carry the disk drives back home. So despite the substantial investments in the internet and the data pipe so to speak, from the perspective of an experiment, it's still somewhat clumsy. So even that infrastructure could do with some attention.

It's also the case that the algorithms that exist have been developed for a fairly specialized set of applications. So, you know, the deep-learning methods, they exist, and what we're doing at the moment is basically borrowing them and applying them everywhere that we can. But, in other words, we haven't gone very far with developing the specialized techniques or the specialized applications.

So even that little movie that I showed, to be honest, I mean, the furthest that we've got is doing very basic analysis so far, and we actually need cleverer, more sophisticated algorithms to analyze all of that information that's latent in those images. I know that sounds like I'm not doing my job, but, I'm just trying to get some idea across of the challenges of taking techniques that have been worked up and then taking them to a completely different domain and doing something worthwhile.

Mr. VEASEY. I was also hoping that you'd be able to describe the progress your group has made in teaching computers to recognize different kinds of metal power—powders using——

Dr. ROLLETT. Powders.

Mr. VEASEY. —additive manufacturing. I think that you——

Dr. ROLLETT. Right.

Mr. VEASEY. —go on to say that these successes have the potential to impact improvements to materials, as well as the generation of new materials. And I hope—was hoping you could talk about that a little bit more and for the ability of a computer to recognize different types of metal and improvements to materials and how that can impact the development of new materials.

Dr. ROLLETT. So thank you for the question. So I was trying to think of a powder—I mean, think of talcum powder or something like that. You spread some on a piece of paper and you look at it and you think, well, that powder looks much like any other powder. It looks like something you would use in the garden or whatever. So the point I'm trying to get across is that when you take these pictures of these materials, one material looks much like another. However, when you take pictures with enough resolution and you allow these machine-learning algorithms to work on them, then what you discover is they can see differences that no human can see.

So it turns out that you can use the computer to distinguish powders from different sources, different materials, so on and so forth. And that's pretty magic. That means that you can again, if you're a company and you're using these powders, you can detect whether you've got—you know, if somebody's giving you what's supposed to be the same powder, you can analyze it and say, no, it's not the same powder after all. So there's considerable power in that.

Another example is things break, they fracture, and you might be surprised, but there's quite a substantial business in analyzing failures. You know, bicycles break and somebody has to absorb the liability. Bridges crack; somebody has to deal with that. Well, that's another case where the people involved look at pictures of these fracture surfaces and they make expert judgments.

So one of the things we're discovering is that we can actually, again, use some of the computer vision techniques to figure out if this fracture is a different kind of fracture or this is a different fatigue failure that's occurred. Again, it's magic. It opens up—not eliminating the expert, not at all. The analogy is with radiography on cancers. It's helping the experts to do a better job, to do a faster job, to be able to help the people that they're working for.

Mr. VEASEY. Thank you very much. I appreciate that.

And, Mr. Chairman, I yield back.

Chairman WEBER. Thank you, sir.

The gentlelady from Arizona is now recognized.

Mrs. LESKO. Thank you, Mr. Chairman.

I have to say this Committee is really interesting. I learn about all types of things and people studying the brains. I think we're going to hear about flying cars sometime soon, which is exciting. I'm from Arizona, and the issues that are really big in my district, which are the suburbs of Phoenix mostly, are actually national security and border security. And we have two border ports of entry connecting Mexico and Arizona, and I have the Luke Air Force Base in my Congressional district. And so I was wondering if you had any ideas how machine learning, artificial intelligence are being used in border security and national security. If you have any thoughts?

Dr. YELICK. Well, I can say generally speaking that in national security, like in science, you're often looking for some signal, some pattern in very noisy data. So whether you're looking at telephones or you're looking at some other kind of collected information, you are looking for patterns. And machine learning is certainly used in that.

I'm not aware in border security of the current applications of machine learning. I would think that things like face-recognition software would probably be useful there, and I just don't know of the current applications.

Dr. NIELSEN. So I know some of the colleagues at our research center are exploring things like security, using facial recognition but trying to take it a step further, so using principles of machine learning, et cetera, trying to detect the intent of a person. So they'll use computer vision, they'll watch a group of individuals but try to infer, make inferences about the intent of what that group is doing. Is there something going to happen? Who is in charge of this group? What are they trying to do?

And they're working with the Department of Defense on many of these applications. And I think there's going to be tremendous breakthroughs where artificial intelligence and machine learning are going to help us not only recognize people but also trying now to recognize the intent of what that person is trying to do.

Dr. ROLLETT. And you mentioned an Air Force Base, so something that maybe not everybody's aware of is that the military operates very old vehicles, and they have to repair and replace a lot. And that means that manufacturing is not just a matter of delivering a new aircraft; it's also a matter of how you keep old aircraft going. I mean, think of the B–52s and how old they are.

And so there are very important defense applications for machine learning, for manufacturing, and manufacturing in the repair-and-replace sense. And again, when you're running old vehicles, you're very concerned about outliers, which hasn't come up very much so far today, but taking data and recognizing where you've got a case that's just not in the cloud, it's not in with everybody else and figuring out what that means and how you're going to deal with it.

Mrs. LESKO. Anyone else? There's one person left.

Dr. KASTHURI. Of course, yes. It's me. So of course my work doesn't deal directly with either border security or national security, but just to echo one other sentiment, one of the things I'm interested in is that, as our cameras get faster, instead of taking 30 shots per second, we can now take 60 shots per second, 90 shots per second, 120 frames per second usually, and you start watching people's facial features as they are just engaging in normal life. It turns out that we produce a lot of microfacial features that happen so fast and so quick that they often aren't detected consciously by each other but convey a tremendous amount of information about things like intent and et cetera.

I suspect that, as our technology, as our cameras get better and of course if you take 120 pictures in a second versus 30 pictures in a second, that's already four times more data that you're collecting per second. If we can deal with the data and get better cameras, we will actually be making inferences about intentions sooner rather than later.

Mrs. LESKO. Very interesting. I'm glad that you all work in these different fields.

And I yield back my time, Mr. Chairman.

Chairman WEBER. Thank you, ma'am.

The gentleman from Illinois, Mr. Foster, is recognized.

Mr. FOSTER. Thank you, Mr. Chairman. And thank you to our witnesses.

And, let's see, I guess I'll start with some hometown cheerleading for Argonne National Lab, which—and I find it quite remarkable. Argonne lab has been—they've come out to events that we've had in my district dealing with the opioid crisis, I find it incredible that one single laboratory—we have everything from using the advanced photon source and its upgrades to directly image what are called G-coupled protein receptors at the very heart of the chemical interaction with the brain all the way up through modeling the high-level function of the brain, the Connectome, and everything in between. And it's really one of the magic things that happens at Argonne and at all of the—particularly the multipurpose laboratories, which are really gems of our country.

Now, one thing I'd like to talk about—and it relates to big data and superconducting—is that you have to make a bunch of technological bets in a situation where the technology is changing really, really rapidly. You know, for example, you have the choice of—for the data pipes, you can do conventional, very wide floating point things for partial differential equations and equations of state, things like that, the way supercomputing has been done for years, and yet there's a lot of movement for artificial intelligence toward much narrower data paths, you know, 8 bits or even less or 1 bit if you're talking about simulating the brain firing or not.

You know, you have questions on the storage where you can have—classically, we have huge external data sets, you know, like the full geometry of the brain that you will then use supercomputing to extract the Connectome. Or now we're seeing more and more internally generated data sets like these are games playing each other where you just generate the data, throw it away. You don't care about storage at all. Or simulation of billions of miles of driving where that data never has to be stored at all, and so that really affects the high-level design of these machines.

In Congress, we have to commit to projects, you know, on a sort of five-year time cycle when every six months there are new disruptive things. We have to decide are these largely going to be front ends to quantum computing or not? And so how do you deal with that sort of, you know, internally in your planning? And should we move more toward the commercial model of move fast, take risks, and break things, or do we have—are our projects that we have to approve in Congress things that have to have no chance of failing? And do you think Congress is too far on one side or the other of that tradeoff?

Dr. YELICK. I guess as a computer scientist maybe I'll start here and I would say that you've asked a very good question. I think this issue of risk and technology is very important, and we do need to take lots of risks and try lots of things, especially right now as not only are processors not getting any faster because of the end of Dennard scaling, but we're facing the end of Moore's law, which is the end of transistors getting denser on a chip. And we really need to try a number of different things, including quantum, neuromorphic computing, and others.

The issue of even the design of computers, if we look at the exascale computing program, very important. Of course, the first machine targeted for Argonne National Lab is in 2021, and the process that is really fundamental to the exascale project is this idea of codesign, that is, bringing together people who understand the applications like Tony and with the people that understand the applied mathematics, and people that understand the computer architecture design.

And the exascale program is looking at both applying machine-learning algorithms for things like the Cancer Initiative, as well as the microbiome where you also have these very tiny datatypes, only four characters that you can store in maybe two bits, and putting all of that together. So those machines are being codesigned to try to understand all those different applications and work well on the traditional high-performance simulation applications, as well as some of these new data-analysis problems.

To answer your question directly, I think that, if anything, that project is very focused on that goal of 2021, and some other machines will come after that in '22 and '23. And the application—so it's not just about delivering the machines; it's about delivering 25 applications that are all being developed at the same time to run on those machines.

It is a very exciting project. I actually lead the microbiome project in exascale, and I think it's a great amount of fun. But it is a project that doesn't have much room for risk or basic research, and so I do think it's very important to rebuild the fundamental research program, for example, the Department of Energy to make sure that ten years from now we could have some other kind of future program that we would have the people that are trained in order to answer those basic questions and figure out how to build another computing device of some kind.

Mr. FOSTER. Well, yes, thank you. That was a very comprehensive answer. But if you could just in my last one second here just sort of—do you think Congress is being too risk-averse in our expectations or, you know, should we be more risk-tolerant that allow you occasionally to fail because you made a technological bet that is—you know, that has not come through?

Dr. YELICK. You know, I think I'll answer that from the science perspective. As a scientist, I absolutely want to be able to take risks and I want to be able to fail. I think the Congressional question I will leave to you to debate.

Mr. FOSTER. Thank you. I yield back.

Chairman WEBER. Thank you.

The gentleman from California, Mr. Rohrabacher, is recognized.

Mr. ROHRABACHER. Thank you very much, Mr. Chairman.

I wanted to get into some basics here. This is for the whole panel. Who's going to be put out of work because of the changes that you see coming as we do what's necessary to fully understand what you're doing scientifically? Who's going to be put out of work?

Dr. ROLLETT. I hope very much that nobody's going to be put out of work.

Mr. ROHRABACHER. Oh, you've got to be kidding. I mean, whenever there's a change for the better, I mean, otherwise, we'd have people working in——

Buggy whips would still be——

Dr. ROLLETT. Yes. I think the point here is to sustain American industry at its most sophisticated and competitive level.

Mr. ROHRABACHER. What professions are going to be losing jobs? You're making me—I mean, everybody's afraid to say that. Come on, you know?

Dr. ROLLETT. I would say they've mostly been lost. I mean, if you look at steel mills, we have steel mills. They used to run with 30,000 people.

Mr. ROHRABACHER. Right.

Dr. ROLLETT. That's why the population of Pittsburgh was so large years ago, right? It's decreased enormously——

Mr. ROHRABACHER. Okay. Well, where can we expect that in the future from this new technology or this new understanding of technology? Anybody want to tell me?

Dr. KASTHURI. I have a very quick——

Mr. ROHRABACHER. Don't be afraid now.

Dr. KASTHURI. I have a very quick answer. Historically, a lot of science is done on getting relatively cheap labor to produce data and to analyze data, by that I mean graduate students, postdoctoral fellows, young assistant professors, et cetera. I suspect——

Mr. ROHRABACHER. So they're not going to be needed probably?

Dr. KASTHURI. Well, I suspect that they should still be trained but then perhaps that they won't be used specifically in just laboriously collecting data and analyzing data.

Mr. ROHRABACHER. Okay. So let's go through that. Where are the new jobs going to be created? What new jobs will be created by the advances that you're advocating and want us to focus some resources on?

Dr. KASTHURI. I'm hoping that when the people who are trained in science no longer have to do all of that work, they do—they then expand into other fields that could use scientific education like the legal system or Congress.

Mr. ROHRABACHER. But what specifically can we look at, say, that will remind Congressmen always to turn off the ringer even when it's their wife? Now, I'm in big trouble, okay? Tell me—so, what jobs are going to be created? What can we expect from what your research is in the future? Do you have a specific job that you can say this—we're going to be able to do this, and thus, people will have a job doing it?

Dr. YELICK. Well, I think there will be a lot more jobs in big data and data analysis and things like that and more interesting jobs I think going along with what was already said, that it's really about replacing—so if we replace taxi drivers with self-driving cars that eliminates a certain class of jobs but it'll——

Mr. ROHRABACHER. Okay. Well, there you go.

Dr. YELICK. Right, but it allows people to then spend their time doing something more interesting such as perhaps analyzing the future of the transportation system and things like that.

Mr. ROHRABACHER. Well, but taxicab driver—finally, I got somebody to admit somebody's going to be hurt and going to have to change their life. And let me just note that happens with every bit of progress. Some people are left out and they have to form new type of lifestyles, and we need to understand that. Maybe we need to prepare for it as we move forward.

What diseases do you think that—especially when we're talking about controlling things that are going on in the human mind, what diseases do you think that we can bring under control that are out of control now? Diabetes, obviously has something to do with the brain is telling the body what to do, different—maybe even cancer? What diseases do you think that we can have a chance of curing with this?

Dr. KASTHURI. I think there's a range of neurological diseases that obviously we'll be able to do a better job curing or ameliorating once we understand the brain. These range from neurodegenerative diseases like Alzheimer's and Parkinson's to more mental illness, psychiatric illnesses and to even early developmental diseases like autism. I think all of these will absolutely be benefited by a better understanding——

Mr. ROHRABACHER. Then if we can control the way the brain is functioning, the maladies that you're suffering like I say diabetes and et cetera, that maybe we can tell the brain not to do that and once we have that deeper understanding.

One last question. I got just a couple seconds. I remember 2001 Hal got out of control and tried to kill these people. And Elon Musk is warning us. I understand somebody's already brought that up. But if we do end up with very independent-minded robots, which is what I think we're talking about here, why shouldn't we think of that as a potential danger, as well as a potential asset? I mean, Elon Musk is right in that.

Dr. ROLLETT. Well, I was going to throw in that I think one opportunity would be in health care and for example, the use of robots as assistants, so not replacing people but having robots help them. Well, those robots have to be programmed, they have to be built.

Mr. ROHRABACHER. Right.

Dr. ROLLETT. I mean, there's a huge infrastructure that we don't have.

Mr. ROHRABACHER. Yes, but if you were building robots that can think independently, who knows—you know, and they're helping us in the hospitals or wherever it is, what if Hal gets out of control?

Dr. ROLLETT. Right, right. So I think AI is being discussed mostly in the context of how do you do something? How do you make something work? When it comes to what these machines actually do, you also need supervision. And what I think we have to do is to build in AI that addresses control and evaluation, you know, the equivalent of the little guy on your shoulder saying don't do that; you're going to get into trouble. So you need something like that, which I haven't heard people talk about much.

Mr. ROHRABACHER. Okay. Well, thank you very much, Mr. Chairman. I yield back.

Chairman WEBER. You've been watching too many Schwarzenegger films.

Mr. ROHRABACHER. That's true.

Chairman WEBER. The gentleman yields back and, Mr. McNerney, you're recognized for five minutes.

Mr. MCNERNEY. I thank the Chairman. And I apologize to the panel for having to step in and out in the hearing so far.

Mr. Nielsen, I'm a former wind engineer. I spent about 20 years in the business. And I understand that the digital twin technology has allowed GE to produce—to increase production by about 20 percent. Is that right?

Dr. NIELSEN. About five percent on an average wind turbine, yes.

Mr. MCNERNEY. Five percent?

Dr. NIELSEN. Five percent, which is pretty amazing when you think we're not switching any of the hardware. It's just making that control system on a wind turbine much smarter using a——

Mr. MCNERNEY. And five percent is believable.

Dr. NIELSEN. Five percent——

Mr. MCNERNEY. Twenty percent for the wind farm——

Dr. NIELSEN. No—yes, it's five percent for——

Mr. MCNERNEY. Okay. Okay. I can believe that. As Chair of the Grid Innovation Caucus, I'm particularly interested in using new technology to create a smarter grid. We have things like the duck curve that are affecting the grid. How can all this technology improve grid stability and reliability and efficiency and so on?

Dr. NIELSEN. Yes, so we're now embarking on research for understanding how to better integrate disparate power sources together in regional, so imagine us trying to use AI machine learning, say, okay, I have a single combined-cycle power plant. How do I better optimize the efficiency of it, produce less emissions, use less fuel, allow more profit from it? But we're taking that now a step further and saying how do I then look regionally and integrating not only that combined-cycle power plant but the solar farm, the wind farm, et cetera? How do I balance that and optimize at a grid-scale level versus just a microscale level?

So that's some of the research that's ongoing now. We're continuing to work on it. But that's our plan is to better figure out that macroscale optimization problem.

Mr. MCNERNEY. So, I mean, once you get that figured out, then you need to have some sort of a SCADA or control system that can dispatch and——

Dr. NIELSEN. Yes, correct.

Mr. MCNERNEY. Okay. So that's another product for GE or for the other——

Dr. NIELSEN. Yes. Correct.

Mr. MCNERNEY. Okay.

Dr. NIELSEN. We're figuring out how to not only build those optimization routines but how to then put them in what we call edge devices, the SCADA systems, the——

Mr. MCNERNEY. Sure.

Dr. NIELSEN. —unit control systems, et cetera. So it's not only trying to figure out the algorithm but making sure that algorithm can execute in a timescale that can be put into some of these, as you mentioned, SCADA systems and control systems.

Mr. MCNERNEY. Okay. Well, with the digital ghost, the—a power plant can replicate an industrial system and the component parts for cyber vulnerability. Is that right?

Dr. NIELSEN. So we use digital ghost at what we call the cyber physical layer. So imagine having a digital twin of a gas turbine. So that digital twin tells us how that gas turbine is behaving and should behave. We then compare to what signal is being generated, what sensors are being—signal's been generated, and we compare that behavior and say that behavior doesn't look right. Our digital twin says something's not correct. The thermodynamics aren't correct.

Mr. MCNERNEY. Well, I mean, I can see that for mechanical——

Dr. NIELSEN. Yes.

Mr. MCNERNEY. —systems. What about cyber?

Dr. NIELSEN. So what we're doing is we're not applying it at sort of the network layer. We're not watching network traffic. We're actually looking at the machine level and understanding if the machine is behaving as it should be given the inputs, the control signals, as well as the outputs, the sensors, et cetera. Some recent attacks look at replicating sensors——

Mr. MCNERNEY. So the same sort of behavior characteristics are going to be monitored—can tell you whether or not there's a cyber issue or some other sort of mechanical failure——

Dr. NIELSEN. Yes.

Mr. MCNERNEY. —impending?

Dr. NIELSEN. Perfect. It's a——

Mr. MCNERNEY. Very good.

Dr. NIELSEN. It's an anomaly detection scheme, yes.

Mr. MCNERNEY. Dr. Yelick, thank you for coming. And I visited your lab a number of times. It's always a pleasure to do so. I think you guys are doing some really good work out there.

One of the things that was striking was the work you did on exascale computing, simulating a San Francisco earthquake and how striking that is. Do you think we have the collective use—have we collectively used this information to harden our systems, to harden our communities against an earthquake, or is that something that is yet to happen?

Dr. YELICK. That's something that is yet to happen. We're just starting to see some of this very detailed information coming from the simulations. And as I mentioned earlier, even bringing in more detailed data into the simulations to give you better geological information about the stability of a certain region or even a certain local area, a city block or whatever, and using that information is not something that is happening yet but obviously should be.

Mr. MCNERNEY. This is sort of a rhetorical question but somebody can answer it if you feel like. I know we hear about the social challenges of digital technology and AI and big data, you know, in terms of job displacement. Does AI tell us anything about that, about how we should respond to this crisis?

Dr. YELICK. I don't know of any studies that have used AI to do that. People do use AI to understand the market, economics, and things like that, and I'm sure that people are using large-scale data analytics of various kinds, and they certainly are to understand changes in jobs and what will happen with them.

It is, by the way, a very active area of discussion within the computer science community about both the ethics, which you heard about I think at previous hearing of AI, but also the issues of replacing jobs.

Mr. MCNERNEY. Sure. Dr. Rollett?

Dr. ROLLETT. If I might jump in, I would encourage you to think about supporting research in policy and even social science to address that issue because AI displacing people is about education, it's about retraining, it's about how people behave. So we scientists are really at sort of the front end of this, but there's a lot of implications that are much broader than what we've talked about this morning.

Mr. MCNERNEY. All right. Thank you. Mr. Chairman, I yield back.

Chairman WEBER. Thank you, sir.

The gentleman from Florida, Dr. Dunn, is recognized.

Mr. DUNN. Thank you very much, Chairman Weber.

And I want to add my thank you to the panel and underscore my personal belief in how important all of your work is. I've visited Dr. Bobby Kasthuri's lab, a great fan of your work and your energy level. Dr. Yelick, we'll be visiting you in the near future, so that'll be fun, too.

I want to focus on the niche in big computing, which is artificial intelligence, and I apologize I missed that hearing earlier, but it was near and dear to my heart.

I think we all see many potential benefits of artificial intelligence, but there are some potential problems, and I think it serves us to face those as we're having this virtual lovefest for artificial intelligence. You know, and we've known this since at least the '60s. I mean, the Isaac Asimov robotic novels and the robotic laws, the Three Laws of Robotics, which I have in my printout, the copies of in case anybody doesn't remember them. I bet this group does.

But what I want to do is—I also, by the way, was looking for guides for artificial intelligence and I came up with the 12 Boy Scout laws, too, so I don't know how that—so I want to offer some quotes and then get some thoughts from you, and these are quotes from people who are recognizably smart people. Stephen Hawking said, "I think the development of artificial intelligence could spell the end of the human race." Elon Musk, quoted several times here, said, "I think we should be very careful about artificial intelligence.

If I were to guess what our biggest existential threat is, it's probably that." Bill Gates responded, "I agree with Elon Musk and I don't understand why people are concerned."

And then finally, Jaan Tallinn, one of the inventors of Skype, said with "strong and artificial intelligence, planning ahead is a better strategy than learning from mistakes." And went on to say, "It really sucks to be the number-two intelligent species on the planet; just ask the gorillas."

So in everybody's handout you have a very brief summary of a series of experiments run at MIT on artificial intelligence. The first one was named Norman, which was an artificial intelligence educated on biased data, not false data but biased data and turned into a deeply sociopathic intelligence. There was another one Tay, which was really just an artificial intelligence Twitterbot, which they turned loose into the internet, and I think it wasn't the intention of the MIT researchers, but people engaged with Tay and tried to provoke it to say racist and inappropriate things, which it did. And there are some other experiments from MIT as well.

So I want to note, like Dr. Kasthuri, I have sons that are more clever than I, but they are not virtual supermen, nor do they operate at the speed of light, so, you know, there's ways of working with them. I'm not so sure about that with artificial intelligence.

My question first, what are the implications of a future where black-box machine learning, the process can't even be interpreted? You know, once it gets several layers in, we can't interpret it. What's the implications today on that to you, Dr. Kasthuri and Dr. Yelick, if I could?

Dr. KASTHURI. Congressman Dunn, thank you for the kind words to start. And I actually suspect there is a reasonable concern that the things that we develop in artificial intelligence are different than the other things like our children because their ability to change is at the speed of computers as opposed to the speed of our own. So I agree that there's legitimate cause for concern.

I suspect that we will have to come up with lessons and safeguards the same way that we've done with every existential crisis: the discovery of nuclear energy, the application to nuclear weapons. As humans, we do have some history of living on the edge and figuring out how to get the benefit of something and keep the risk at bay.

You're right that if algorithms can change faster than we can think, our existing previous historical safeguards might not work.

To the specific question that you asked about the non-interpretability, for me, without knowing what the algorithm is producing, how do you innovate? If you don't know the fundamental nature of what the algorithm is—its principles for how it comes to a conclusion, I worry that we won't be able to innovate on those results.

And this is interestingly perhaps as a thought exercise: What if a machine-learning algorithm could tell me—could make—could collect enough data to make a prediction about a brain, about your brain or someone else's brain that was incredibly accurate? Would we at that moment care how that machine-learning algorithm arrived at its conclusion? Or would we at that moment take the results that the algorithm produces and just go on with it, in which case there could be a missed opportunity for learning something deeply fundamental and principled about the brain.

Mr. DUNN. And very quickly, Dr. Yelick.

Dr. YELICK. Well, I agree with that. I think that these deep learning algorithms which have these multiple layers, which is why they're deep, they have millions perhaps of parameters inside of them. And we don't really understand when you get an answer out why all these parameters put together tell you that that's a cat and this one's not a cat. And so that may be okay if we're trying to figure out where to place ads as long as we give it unbiased data about where the place the ads so the right—so——

Mr. DUNN. But it might be more problem if it was flying a drone swarm on attack some place?

Dr. YELICK. Well, where it's a problem is if I'm a scientist, I want to understand why. It's not enough to say there's a correlation between these two things. And if the, you know, drone is flying in the right place, that's really probably the most important thing about some kind of a controlled vehicle. But in science, you want to——

Mr. DUNN. We're dangerously close to being way, way, way over time, so I better yield back here, Mr.—thank you very much, though. I appreciate the chance.

Chairman WEBER. All right. The gentlelady from Nevada, Ms. Rosen, is recognized.

Ms. ROSEN. Thank you. I want to thank you for one of the most interesting, informative, and I want to say this is on the bleeding edge of everything that we need to worry about for sure.

But one thing we haven't talked about is data storage. And data storage specifically is critical infrastructure in this country, right, because we have tons and tons of data everywhere, and where it goes and how we keep it is going to be of utmost importance.

And so I know that we're trying to focus on that in the future, and in my district in Nevada we have a major data storage company. It has state-of-the-art reliability. We have lots of quality standards to ensure its data is secure, but like I said, we don't consider it critical infrastructure.

So right now in this era of unprecedented data breaches, data hacks, every moment they are just pounding on us, in your view what are—the data storage centers that house the government and private sector, where are their vulnerabilities and what are the implications? How should we be sure that we classify them as critical infrastructure?

Dr. YELICK. So, clearly, those data centers are storing very important information that should be protected. And, as you said, even at the computing centers that we run in the labs, there's a constant barrage of attacks, although we store at NERSC the center at Berkeley lab only scientific data, so it is not really critical data. I think that using these kinds of machine-learning techniques to look for patterns is one of the best mechanisms we have to prevent attack, and they do have to learn from these patterns in order to figure out what is—and—what is abnormal behavior. And we're looking at—as we build out the next network, even kind of embedding that information into the network so that you can see patterns of attack even before they get to a particular data set or a particular computer system.

Ms. ROSEN. Thank you. I have one other question. And you were talking about using predictive analytics with a digital twin to talk about fatigue in planes. But how can we use that to discuss infrastructure fatigue as we talk about the infrastructure failures around this country in bridges, roads, ports, et cetera, et cetera? So——

Dr. ROLLETT. That's I think a question of recognizing the need and talking to the agencies and finding out whether you consider there are adequate programs to do that. I'm going to guess that there is not a huge amount of activity, but I don't know, so that's why I'm being very cautious in my answer.

But I suspect it's one of the opportunity areas. It's an area where there is data. It's often rather incomplete, but it would definitely benefit from having the techniques applied, the machine-learning techniques to try to find the patterns, to try to identify outliers, particularly trends that are not good.

Ms. ROSEN. Thank you.

Dr. NIELSEN. I would just——

Ms. ROSEN. Oh, please, yes. Yes.

Dr. NIELSEN. Oh, I'm sorry. I would just second the comments made. I mean, at GE we obviously focus a lot of our attention on the commercial assets that we build, but there's no reason the technologies, the ideas that are being applied there could be applied to bridges and infrastructure and all that.

Ms. ROSEN. Right.

Dr. NIELSEN. It's just, I think, a matter of will and policy to do that, right?

Ms. ROSEN. So I—do you think that would be well worth our time here in this Committee to promote those kinds of policies or research for you all or someone to do the—use the predictive analytics? Congresswoman Esty and I sit on some infrastructure committees, and really important that we try to find out points of failure before they fail, right?

Dr. ROLLETT. Absolutely. And I would encourage you to bring state and local government into that discussion because they often own a lot of those assets.

Ms. ROSEN. Yes. Thank you. I yield back my time.

Chairman WEBER. The gentlelady yields back.

The gentlelady from Connecticut is recognized.

Ms. ESTY. Thank you so much. And this is tremendously important for this Committee and for the U.S. Congress to be dealing with, and we really appreciate you taking the time with us today.

All of you have mentioned somewhat in passing this critical importance of how are the algorithms structured and how are we going to embed the values if we have AI moving much faster than our brains can function or at least on multiple levels simultaneously?

So we did have a hearing last month in talking about this, and one of the issues that came up that everyone supported—and I'd like your thoughts on that—is the critical importance of a diverse workforce in doing that. If you're going to try to train AI, it needs to represent the diversity of human experience, and therefore, it can't be like my son who did computer science in astrophysics. If they all look like that, if those are—the algorithms are all being developed by, you know, 26-year-olds like my son Thomas, we're not going to have the diversity of life experience.

So, first, if you can quickly—because I've got a couple of questions—thoughts on how do we ensure that? Because we're looking at that issue. We talk about that diverse workforce all the time, but when we're looking at AI and algorithms, it becomes vitally important that we do this. It's not about checking the box to say the Department of Labor that we've got a diverse workforce. This is actually vital to what we need to do.

Dr. YELICK. So if I can just comment on that. Yesterday, before I left UC Berkeley, I gave a lecture to the freshman summer class introductory computing class. My title was rather ostentatious as "How to Save the World with Computing." What I find is that when you talk about the applications of computing and including data analytics and machine learning and real problems that are societal problems, you tend to bring in a much more diverse workforce. That class in particular has had over 50 percent women and a very good representation at least relative to the norm of underrepresented minorities as well.

Ms. ESTY. Anyone else who—I mean it—MIT has found that when they change the title of some of their computer science classes to again be applied in sort of more political and social realms, they had a dramatic change in terms of composition of classes.

Dr. NIELSEN. Yes, I would just quickly build upon that, too. I think to me when you look at AI and machine learning, you have to have a critical eye. You have to always be looking at it. And I think a diverse workforce and diverse experience can help just bring more perspectives to help critically question why are those algorithms doing what they're doing? What is the outcomes? How can we improve that? So I would support that supposition, yes.

Dr. YELICK. I'll just mention that the name of the course—which I was not teaching, by the way, I was giving a guest lecture—is "The Beauty and Joy of Computing," so maybe that helps.

Ms. ESTY. Well, that helps. And if I could have you turn again—and some of you have mentioned the important role of federal research. I mean that's what this Committee is looking at, what is uniquely the federal role. As you see across the board, there's more and more effort and being engaged and we see it in space research and other places to move into the private sector with the notion the federal government is not very good at picking winners and losers. So if you can all talk about what you think are the most critical tasks for federal investment in, say, foundational and basic research that then will be developed by the GE's and others and companies not yet formed or conceived of because, again, that's part of our job is to figure out—I see it as our job to defend putting those basic research dollars in because we don't know where they're going to go but we do know they're vital to keep us, whether it's competitive or frankly just have better research and more care.

Dr. KASTHURI. So perhaps I can go really quick. I suspect that there is a model of funding scientific research that's this idea that if you plant a million seeds in the ground, a few flowers will grow, where individual labs and individual scientists have the freedom to judge what is the next important question to address.

And I can see why having the federal government decide the next important question to address might not be the most efficient way to push science forward. But where I do see the federal government really playing a role is in the level of facilities and resources, that what I imagine is that the federal government establishes large-scale resources and facilities like the national lab system and then allow individual scientists to promote their individual ideas but leveraging the federal resources. And I wonder if this is a compromise between allowing these seeds to grow but the federal government—maybe this is appropriate but maybe not—providing the fertilizer for those seeds.

Ms. ESTY. They think we generate a lot of it at least in this place.

Dr. YELICK. So I would just add I think the importance of fundamental research, as well as the facilities and infrastructure and the applied mathematics, the computer science, statistics, very important in machine learning. And, as we said, these machine-learning algorithms have been used a lot in nonscientific domains. There's a lot of interest in applying them in scientific domains. I think the peer-review process in science will make machine learning better for everybody if we really put a lot of scrutiny on it.

Dr. ROLLETT. And very quickly, I wanted to add that I think it's important that program managers in the federal government have some discretion over what they fund and take risks. And it's also important that the agencies have effective means of getting community input. And I don't want to name names, but some agencies have far more effective mechanisms for that than others.

Ms. ESTY. Well, we might want to follow up with that last point.

And I wanted to just put out for you to help us with—and you mentioned it, Dr. Yelick, with—on peer review, this systematic—because of pressures to publish or perish and show success is we are not sharing the failures, which are absolutely essential for science to make progress. It's one of the issues we've touched on a lot in this Committee. We don't have any good answers, and it's gotten worse because of the pressures to do—to get grant money and to show progress. But I am deeply concerned about those pressures both from the private sector and the public sector making it harder for us—people hoard the, quote, "bad results," but they're absolutely essential for us to learn from them.

And so I don't know how we change that dynamic, but I think that is something that we could really use your thoughts on that because whether it's—AI can maybe help us with disclosing the dead ends and we learn from the dead ends and we move forward. But it is something that we have a big issue with in how we deal with the sharing of the not-useful results, which may turn out to be very useful down the line.

Dr. YELICK. I completely agree with that. I think the first step in that is sharing the scientific data and allowing people to reproduce the successful results but also, as you said, examine the supposed failures to see—there are many examples of this in physics and other disciplines where people go back to data that may be 10 or 20 years old and find some new discovery in it.

Ms. ESTY. Thank you very much. I really appreciate your indulgence to keep us here to the bitter end. Thank you. Not the bitter, not you, just the fact that the bell has rung, and we had a lot of questions for you. We appreciate it. Thank you so much.

Chairman WEBER. After failing 1,000 times for the lightbulb, Dr. Edison, his staffer said doesn't that frustrate you? He goes, what are you talking about? We're 1,000 ways closer to success.

So I thank the witnesses for their testimony and the Members for their questions. The record will remain open for two weeks for additional written comments and written questions from the Members.

This hearing is adjourned.

[Whereupon, at 12:08 p.m., the Subcommittees were adjourned.]

Appendix I

ANSWERS TO POST-HEARING QUESTIONS

ANSWERS TO POST-HEARING QUESTIONS

Responses by Dr. Bobby Kasthuri

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Bobby Kasthuri, Researcher, Argonne National Laboratory; Assistant Professor, The University of Chicago

Questions submitted by Rep. Gary Palmer, House Committee on Science, Space, and Technology

1. **In addition to serving here on the Energy Subcommittee, I also serve on the House Budget Committee where I am familiar with looking at government spending from a cost-benefit viewpoint. The President's budget request for FY 2019 included $899 million for the Advanced Scientific Computing Research Program, so I am wondering if you could speak a little more to the benefits that you see coming directly from that investment? In your opinion, what areas/technologies are giving us the best return on investment?**

The United States benefits considerably from U.S. Department of Energy (DOE) investments in the Advanced Scientific Computing Research (ASCR) program. DOE investments have created a modern scientific computing environment that supports research throughout the country and expedites breakthroughs in fields ranging from physics, chemistry, biology, and materials to cosmology, environmental science, energy sciences, and transportation.

Software deployed by thousands of supercomputing users every day to simulate physical, chemical, and biological systems comes from DOE investment—in software development as well as in studies of the fundamental mathematics underpinning the software. DOE contributions enable developers to write code for millions of processors and to move petabytes (a petabyte equals one million gigabytes) of data from coast to coast.

Investments that empower us to address "grand challenges" on the largest supercomputing systems offer the best return on investment because we integrate necessary hardware, software, and mathematics into unified working applications. Industry and academia then copy and build upon these exemplar applications, resulting in returns to the U.S. economy at rates of 100:1 and even 1000:1.

2. **We all know that China is a major competitor in the machine-learning/AI space. What would it mean for the United States if another country were to gain dominance in machine learning?**

The nation that leads in machine learning and artificial intelligence (AI) will lead the world in developing new technologies, medicines, industries, and military capabilities. Most of the modeling and prediction necessary to produce the next generation of breakthroughs in science, energy, medicine, and national security will come not from applying traditional theory, but from employing data-driven methods. Losing dominance in machine learning and AI will result in the United States playing catch-up in dozens of important areas.

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Bobby Kasthuri, Researcher, Argonne National Laboratory; Assistant Professor, The University of Chicago

Questions submitted by Rep. Jacky Rosen, House Committee on Science, Space, and Technology

As you all have discussed, scientists and companies continue to utilize big data analytics to better achieve research goals and improve industry needs. In Nevada, the Desert Research Institute – or DRI, the state's environmental research facility – uses extensive monitoring and modeling programs to analyze environmental and health data. For the past two years, DRI has been working with partners on the Healthy Nevada Project, one of the first community-based population health studies in the country. They are studying health, environmental, and socioeconomic data to better understand how these factors and genetics can help predict who may be at risk for certain diseases, allow for quicker diagnoses, and encourage the development of better treatments.

1. **In your view, is there a productive role that the Department of Energy can play in accelerating the development of technologies like those that the Healthy Nevada Project is using?**

The U.S. Department of Energy (DOE) brings the power of supercomputing and advanced mathematics to challenging issues in healthcare, complementing the capabilities of other federal institutions. Via its joint projects with the National Cancer Institute (NCI) and the U.S. Department of Veterans Affairs (VA), DOE is developing technologies to conduct large-scale genomics analysis and to process medical records with artificial intelligence (AI). These technologies are designed to increase cancer treatment options, and to improve veterans' healthcare through the Million Veterans Program (MVP) and its focus on cardiovascular disease, suicide prevention, traumatic brain injury, and prostate cancer. Researchers could apply the same advanced computing methods to the types of data the Healthy Nevada Project is currently collecting and analyzing.

2. **How should we balance investments in DOE's computing facilities and advanced data analytics? Are we creating data faster than we can analyze it?**

DOE has invested significantly in data analytics over the last five years, focusing on mathematics for data analysis, machine learning, and AI methods to identify patterns in data and build predictive models from those data. DOE also is investing in the necessary software tools and infrastructure to manage extreme data flows from large-scale instruments shared by international scientists, such as the Large Hadron Collider (LHC) in Switzerland. DOE's Fermi National Accelerator Laboratory hosts a Tier-1 computing center that processes data from LHC experiments. Other large-scale, collaborative instruments yielding extreme data flows include light sources such as the Advanced Photon Source (APS) at Argonne National Laboratory and detectors from telescopes and microscopes.

DOE also is investing in the design of computer systems optimized to process data and to run simulations. The Aurora 21 system—which Argonne will deploy in 2021 as the first U.S. exascale system—is explicitly designed to support exascale simulations, the largest data analytics problems, and deep learning for DOE science and engineering missions. It is important for future DOE computing facilities to embrace the convergence of simulation, data analysis, and machine learning, and for DOE to invest in both facilities and research to gain maximum insights from the data collected. With the deployment of exascale computers, DOE will have some of the world's most powerful data-analysis engines, designed to keep pace with the volume of data scientists are creating.

3. **How should the need to accelerate big data analytics and integrate private sector approaches influence the design requirements and success metrics of upcoming DOE computing acquisitions?**

DOE computer scientists are well aware of data analysis approaches developed by the private sector—and in some cases, they have contributed to that technology. Insights from private-sector systems and similar government-developed systems have influenced the design of DOE's preexascale systems (Summit and Sierra), as well as DOE exascale systems that are currently under development (Aurora 21, Frontier, and El Capitan). The need to address simulation, data analytics, and machine learning has resulted in machine architectures and software environments that combine the best of scientific computing and big data analysis. With supercomputing vendors that understand this need for convergence, the United States is the undisputed thought leader in next-generation computing systems.

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Bobby Kasthuri, Researcher, Argonne National Laboratory; Assistant Professor, The University of Chicago

Questions submitted by Rep. Randy Hultgren, House Committee on Science, Space, and Technology

1. **Clearly AI is important to the future of science, but do you feel that our nation's computing infrastructure is equipped with the right kind of computers to enable researchers across our nation to develop this capability?**

Pre-exascale and exascale systems under development at U.S. Department of Energy (DOE) laboratories will be the world's most powerful for artificial intelligence (AI). National laboratory and academic researchers can access these systems through successful peer-review processes the laboratories have used for more than a decade. However, although DOE is establishing the right types of systems in its national laboratories, U.S. academic groups may not have sufficient capacity to teach next-generation researchers how to apply advanced AI methods to science. Consequently, it may make sense for DOE to work with the National Science Foundation (NSF) or others to broaden academic access to leading-edge AI systems via testbeds and facilities housed at the national laboratories. When used in the 1980s and 1990s, this approach exposed academics to the earliest parallel computing systems and grew a research community around parallel processing. Reviving the approach with a focus on AI would complement the resources available to universities via public computing clouds.

2. **How are we including academia and other research institutions with work at DOE?**

The DOE Office of Science supports broad research programs at universities; depending on the DOE program office, academia-based research can constitute as much as 30% of the allocated funding. University-supported researchers have full access to the DOE facilities they need to carry out their work. In addition, DOE gives academic researchers access to facilities at DOE laboratories independent of their funding sources. Many users of DOE supercomputers and light sources are academics supported via NSF, National Institutes of Health, or U.S. Department of Defense research programs. DOE also provides facility and technology access to researchers from non-profit institutes and industrial organizations.

3. **Do you believe we have an AI infrastructure gap and what would a roadmap to getting where we need look like?**

There is an emerging need for the United States to enable wider access to the types of specialized artificial intelligence (AI) computers being developed by startups. Dozens of companies are investing a total of more than $3 billion to push performance boundaries for AI-based applications. Although DOE is involved with some of these projects, overall the United States has been more successful in research and development for these systems than in deploying them for research use.

Although the United States and China currently have roughly equal access to AI infrastructure, China is making serious investments. Japan currently lags compared to the United States, but it has recently started building AI supercomputers to provide access for researchers in academia and industry. In the United States, most AI infrastructure is operated by companies such as Google, Amazon, and Microsoft, whose primary goals do not include making their leading-edge capabilities available to research programs.

The United States could consider deploying at DOE national laboratories AI access infrastructure that offers leading-edge technologies and is available to U.S. universities and smaller industry. As mentioned before, the United States successfully followed this strategy in the 1980s and 1990s to provide academics necessary access to emerging parallel computers.

Responses by Dr. Katherine Yelick

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Katherine Yelick, Associate Laboratory Director for Computing Sciences, Lawrence Berkeley National Laboratory; Professor, The University of California, Berkeley

Questions submitted by Rep. Gary Palmer, House Committee on Science, Space, and Technology

1. **In addition to serving here on the Energy Subcommittee, I also serve on the House Budget Committee where I am familiar with looking at government spending from a cost-benefit viewpoint. The President's budget request for FY 2019 included $899 million for the Advanced Scientific Computing Research Program, so I am wondering if you could speak a little more to the benefits that you see coming directly from that investment? In your opinion, what areas/technologies are giving us the best return on investment?**

The Advanced Scientific Computing Research (ASCR) Program is the lead agency for high performance computing (HPC) in the nation, with the top HPC facilities and a research program in applied mathematics, computer science, and its signature SciDAC Partnerhip program to develop HPC applications in collaboration with other science communities. ASCR's ESnet network connects all the sites together, allowing data to stream from one user facility to another, providing both a unique capability for scientists and greater overall efficiency in science, since one can use supercomputers to analyze data from major experimental facilities. ASCR's HPC facilities are a tremendous resource to the national user community, with the Leadership Computing Facilities supporting some of the largest national computational challenges, and NERSC supporting the broader DOE user community in high performance computing and data analysis. NERSC has a distinct role and ROI, with over 7000 users of which 60% (over 4000) are students, postdoctoral researchers, and faculty from universities. These facilities provide computing and data services by installing ready-to-use application software in a broad range of research areas, in addition to training and support for those who want to develop their own applications. Finally, the Exascale Computing Project (ECP) is a novel construct that has marshaled efforts across the DOE complex to build applications and software for exascale systems, in addition to advanced technology investments in industry. ECP is leveraging basic research investments from the last decade, such as novel mathematical models fast parallel algorithms from computer science in order to solve problems that would otherwise have been impossible.

The combination of these ASCR-funded activities is serving the national scientific community while also moving the nation forward in advanced hardware, software, and mathematical capabilities. Advanced high performance computing (HPC) is increasingly becoming an indispensable and foundational part technologies important in US industry. Today, advanced HPC is speeding the discovery of new materials for lighter yet stronger airplane wings and faster planes, of new chemistry for more efficient batteries, and of larger data analyses for more detailed understanding of extreme weather events and how to plan for and recover from them.

Because of advances in data-driven science funded by ASCR, HPC can now process huge amounts of data to discover interesting features and infer properties of complex scientific data – adding powerful new tools in analysis and learning to established areas of HPC modeling and simulation. The amounts of data available to researchers are increasing exponentially and the United States' ability to understand and productively utilize the information relies on advanced computing capabilities and capacity – from more powerful hardware to more sophisticated mathematical algorithms and more complex, speedier, and efficient software, memory, and networking systems.

Advanced HPC is fundamental to securing the United States' world leadership in science, but it is also critical to advancing the nation's leadership in the production of information technology and products, from the intellectual property of microelectronics, software and advanced applied mathematics, to maintaining world leading market share in computing and communications devices and tools. New knowledge that drives transformational leaps in advanced high performance computing also pushes against the technological boundaries of personal computing devices, internet technologies and other consumer, computing based, products.

2. **We all know that China is a major competitor in the machine-learning/AI space. What would it mean for the United States if another country were to gain dominance in machine learning?**

It is imperative to U.S. scientific, innovation and economic wellbeing that our researchers, universities, institutions, industries and government maintain a world leadership position in the development and utilization of machine learning and AI. In the largest scientific problems, we are not trying to develop techniques that mimic human behavior, but instead augment human insight with the ability to analyze data sets at a scale and complexity that would be impossible for humans. The US still holds a lead in this area, but countries around the world have recognized the on the benefits that AI and machine learning bring to a very wide variety of scientific, national and commercial applications, and are investing heavily in these fields.

Instead of focusing exclusively on AI and machine learning leadership, it is critical to ask how the federal government can work with universities, national labs and industry to support the whole ecosystem of high performance computing and meet the challenges of a beyond Moore's law world. From advanced hardware and high speed scientific networking, to world leading applied math capabilities, advanced memory systems, sophisticated software, and an educated and well-trained workforce in computing and computation, the entire ecosystem must be sustained. Adequately and wisely investing in the foundational pieces of the HPC ecosystem will drive advances, and leadership, in AI and machine learning.

As widely reported a decade ago, China made a commitment to develop an indigenous high performance computing industry that would be able to compete toe to toe in international markets against the U.S., Europe and China – primarily the U.S. The Chinese approach has been broad and successful, and the country's HPC industry is now poised, and in some instances have begun, to harvest the fruits of their investments. In 2011, China claimed 61 out of the top 500 supercomputing sites in the world — making it a distant second to the U.S., which had 255. Today, in the most recent issue of the HPC Top500 list, the U.S. claims only 124 systems, a new low. Just six months ago, the US had 145 systems. Meanwhile, China improved its representation to 206 total systems, compared to 202 on the last list.

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Katherine Yelick, Associate Laboratory Director for Computing Sciences, Lawrence Berkeley National Laboratory; Professor, The University of California, Berkeley

Questions submitted by Rep. Jacky Rosen, House Committee on Science, Space, and Technology

As you all have discussed, scientists and companies continue to utilize big data analytics to better achieve research goals and improve industry needs. In Nevada, the Desert Research Institute – or DRI, the state's environmental research facility – uses extensive monitoring and modeling programs to analyze environmental and health data. For the past two years, DRI has been working with partners on the Healthy Nevada Project, one of the first community-based population health studies in the country. They are studying health, environmental, and socioeconomic data to better understand how these factors and genetics can help predict who may be at risk for certain diseases, allow for quicker diagnoses, and encourage the development of better treatments.

1. **In your view, is there a productive role that the Department of Energy can play in accelerating the development of technologies like those that the Healthy Nevada Project is using?**

Although not within its core mission, health research has and can benefit greatly from the expertise and unique scientific resources of the Department of Energy. The best-known example of this is the Department's role in sequencing the human genome. High performance computing capabilities and the ability to manage large scale scientific research challenges made the Department an ideal place to take on this huge challenge. Today, partnerships between the DOE and the NHI and Veterans Affairs are working to apply unique DOE assets to critical health research into brain science and cancer.

So, yes, there probably is a productive role for DOE to play in assisting the Healthy Nevada Project and other research programs like it. The question is how to do it. Stovepipes within the Congress and the Administration often serve as barriers to cross-agency collaboration. The biggest issues often being who pays. If however, the resources for the research are provided by external sources (foundations, companies, other state, regional, or local governments), the national laboratories have several different points of engagement possible to conduct research for non-DOE entities. Additionally, the DOE national scientific user facilities are free to use based on an external peer reviewed process that ranks research proposals based on the quality of the science and the appropriateness for the specific user facility.

2. **How should we balance investments in DOE's computing facilities and advanced data analytics? Are we creating data faster than we can analyze it?**

In some cases, yes, DOE's (and other agencies') data generating tools and experiments, such as light sources, telescopes, sensors, accelerators, genome sequencers particle colliders, etc., are producing data faster than ever before. However, this data will produce new insights using advanced analysis techniques and systems optimized for data-intensive workloads. We understand the challenge and the Department has a plan in place to grow the nation's computing capabilities to handle massive data opportunities. However, the challenge is not two dimensional. It is not just about computing power and data production. It is about the entire HPC and scientific computing ecosystem. The data needs to move to the HPC facilities over high speed networks and analyzed with smarter algorithms, including machine learning techniques, and scalable parallel versions. More sophisticated and flexible software, memory and connectivity are required. Advanced scientific networking between the sources of data production and the HPC assets must be able to seamlessly and accurately move the scientific data. Because of this, it is critical that DOE's HPC budget remain balanced among the different foundational parts of the ecosystem. Without this balance, the U.S. will continually be playing catchup.

3. **How should the need to accelerate big data analytics and integrate private sector approaches influence the design requirements and success metrics of upcoming DOE computing acquisitions?**

DOE's computing and networking facilities have well-established processes for collecting scientific requirements from the user community and translating them into system requirements for acquisitions. At the National Energy Research Scientific Computing Center (NERSC), for example, data analytics has been part of the workload for many years, and one part of the current Cori systems is tailored to the need of some data-intensive problems. NERSC used to run separate systems for data analysis, but having single integrated systems are more flexible and more cost effective. At the same time, the private sector, both large companies and small startups, are developing innovative hardware approaches for some of these problems, especially for deep learning algorithms. The DOE Labs would bring a wealth of expertise in algorithms and scalable software, as well as large, complex scientific data sets to aid in co-design of these systems.

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Katherine Yelick, Associate Laboratory Director for Computing Sciences, Lawrence Berkeley National Laboratory; Professor, The University of California, Berkeley

Questions submitted by Rep. Randy Hultgren, House Committee on Science, Space, and Technology

1. **Clearly AI is important to the future of science, but do you feel that our nation's computing infrastructure is equipped with the right kind of computers to enable researchers across our nation to develop this capability?**

The Deep Learning approaches that have been successful in many AI problems require enormous computing resources and are dominated by dense matrix algorithms that run well on most computer architectures, but are especially efficient on Graphics Processing Units (GPUs). With Large GPU-based systems on the floor--such as the recently installed Summit system at ORNL and Sierra at LLNL, in addition to the older Titan system at ORNL--DOE is providing capabilities that are well matched to deep learning. The DOE systems have high-speed networks, which allow even large data sets and enable more scalable algorithmic approaches than on smaller systems or those with slower networks, as are common in commercial clouds. Indeed, the largest and fastest deep learning problems run to date have been on these DOE systems. Just a year ago the NERSC Cori system with its lightweight cores (not GPUs) produced the fastest deep learning demonstration used to classify scientific data, and a demonstration of one exaops will be reported at the upcoming Supercomputing (SC18) conference by a team, lead by Berkeley Lab on the Summit system. (These "ops" used in AI problems are one quarter the power of our usual "flops" we talk about in HPC, so you can roughly think of this as a 250 petaflop system.) Longer term, there are several companies and academic researchers looking at even better architectures for these problems, which may be able to solve problems faster or with less energy. The DOE Lab's close partnerships with many of these vendors, our experience pushing the envelope of new technologies, and the availability of enormous scientific data set and expertise can contribute to advances in hardware for machine learning.

2. **How are we including academia and other research institutions with work at DOE?**

Across the Office of Science, universities receive substantial research funding, both a direct awards to universities and through subcontracts and partnerships with the DOE Labs. They also gain access to DOE user facilities, which enable research that would otherwise be impossible at most university campuses. With respect to computing, as noted above, over 4000 university users benefit from using NERSC, which includes scientific user support in addition to access to computing and data systems. ESnet impacts universities in multiple ways, peering with other networks to move scientific data between DOE facilities and the universities, and providing innovative services and architectures, such as the Science DMZ architecture that NSF adopted and funded at several universities in order to speed science data transfers.

Academic institutions are involved in some projects within the Exascale Computing Program (ECP) and they are directly funded by ASCR's base research program. However, funding for base computer science and applied math research has declined significantly over the past few years as the ECP project ramped up. The research program needs to be rebuilt to lay the foundation in ideas, research results, and in personnel for future mission needs within DOE. DOE needs to have longer-term, high risk research for the long-term health of the Labs and to attract the best talent, especially in HPC where there are so many other opportunities for computational experts entering the workforce. While ECP is serving an important role in the next few years, it should not replace basic research. University-funded researchers have been even more significantly impacted by reductions in the base research program than the Labs, which, of course, is a critical issue for current students and faculty, but has negatively impacted both the innovation and talent pipeline needed to address future DOE mission problems and computational / data-intensive science more broadly.

3. **Do you believe we have an AI infrastructure gap and what would a roadmap to getting where we need look like?**

Ironically, most computer scientists have not used a large amount of computing time for their own research, unless they were working (as I have been) on HPC problems in partnerships with scientists from other disciplines. But the explosion of machine learning, and especially deep learning, in AI has suddenly created a demand for access to large computational resources by computer scientists. While DOE will need machine learning to addressing some of its mission problem, and all of the DOE HPC facilities are looking at way of supporting that workload, planned upgrades to these facilities will not serve the needs of this computer science community, many of whom are working on problems outside the DOE mission space. The National Science Foundation also provides HPC resources, but this has primarily focused on the physical and life sciences, not on computer science. Commercial cloud providers also offer computing services and even provide modest sized academic "grants" of cloud time, but costs can be prohibitive if university researchers are paying commercial prices and these systems do not provide the high performance computing at scale needed for some of the largest problems. As a result, academic computer scientists do not have access to the kind of resources (both data and computing) that are available to industry researchers. This has created a gap that is threatening the intellectual leadership of universities in the AI space. While there may be some differences in architectural designs, such as optimizations of hardware or software specific to deep learning, the real issue is simply lack of academic infrastructure for machine learning, because the need has grown so quickly and has not been met by existing facilities. DOE's expertise in running large facilities for the science community, including innovative computer systems, collecting user requirements, installing software, and providing scientific support, makes it well-positioned to provide some of the computing need to AI.

Responses by Dr. Matthew Nielsen

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Matthew Nielsen, Principal Scientist, Industrial Outcomes Optimization, GE Global Research

Questions submitted by Rep. Gary Palmer, House Committee on Science, Space, and Technology

1. **In addition to serving here on the Energy Subcommittee, I also serve on the House Budget Committee where I am familiar with looking at government spending from a cost-benefit viewpoint. The President's budget request for FY 2019 included $899 million for the Advanced Scientific Computing Research Program, so I am wondering if you could speak a little more to the benefits that you see coming directly from that investment? In your opinion, what areas/technologies are giving us the best return on investment?**

Three significant benefits to industry that the Advanced Scientific Computing Research Program (ASCR) delivers are (1) the ability to evaluate computational solutions to barrier problems beyond what can be feasibly tested internally, (2) collaboration with experts in computational methods to push beyond the state of the art and (3) access to leverage the ecosystem of software tools and applications to benefit from advances in employing computational solutions. The ASCR program offers access via competitive peer-reviewed grants to time on supercomputers with capabilities beyond any feasible investment from industry. As access to these grants has broadened to allow participation from industry, GE has competed aggressively through those programs to earn supercomputing time – and access to the critically-valuable talent at the computing centers that comes with such a win. Harnessing the power of these facilities and the knowledge of the technical staff, GE has gained insight into scientific and engineering problems and solutions with impact on global competitiveness in advanced technology products ranging from power generation to jet propulsion to metal 3D printing to medical imaging. In fact, expertly employing computational methods is now fundamental in the modern practice of engineering design.

ASCR's leadership computing facilities assemble the investments, talents, skills and knowledge to: (a) advance technology, procure and operate the most powerful computers in the world (beyond the horizon pragmatic for industry to provide on its own) (b) develop the software to maximize the usefulness and broaden the applicability of leveraging these facilities for a multitude of problems critical to domestic interests and global competitiveness and (c) coordinate such efforts in exemplary programs such as the flagship Exascale Computing Project (ECP), for which GE presently Chairs the Industry Advisory Council. For example, the ECP includes initiatives to propel cutting edge capability in additive manufacturing, combustion machinery and cancer treatment.

Further, newly-deployed systems such as Summit at the Oak Ridge Leadership Computing Facility recognize the emergent power of human+machine collaboration – applying the strengths of each as machine learning and scalable data analytics augment human perception, cognition and comprehension to overcome challenges in ever-more volatile, uncertain, complex and ambiguous in science, medicine, engineering and economics. Recent government support of quantum computing demonstrates commitment to longer-term impacts, and already quantum communication and cryptography are beginning to find commercial applicability. Quantum computing remains more distant in its readiness to be used on the very narrow set of problems to which it is known to be applicable. In those niche areas, quantum computing is truly transformative and there will be national consequences if the U.S. falls behind. However, the value of advanced-but-conventional computing programs such as ECP is established, significant and widely-applicable and should not be interpreted in any way as redundant to or inevitably replaced by quantum computing.

2. **We all know that China is a major competitor in the machine-learning/AI space. What would it mean for the United States if another country were to gain dominance in machine learning?**

The United States must become a leader in artificial intelligence to preserve our national security and maintain global peace. Any nation that gains artificial intelligence dominance over the United States has the option to defeat us across several theaters, including but not limited to: cyber, air, sea, ground, space, economic, biologic, and to destroy us from within through inciting social conflict.

Our nation has an economy that depends on the internet, just as much as we depend on electricity, water, food and shelter. The cyber war is currently being waged, but we have not experienced a catastrophic cyber event that signals to the world that one country has superior cyber power. Machine Learning experts that have both created and witnessed the power of AI Systems that continuously learn and they understand the potential that artificial intelligence brings to cyber warfare. The United States needs to invest at the intersection of Artificial Intelligence and Cyber Security with the domain knowledge of our assets that compose our critical infrastructure that deliver power, water, food, transportation and healthcare – within the context of our digital world.

An upper hand in Artificial Intelligence is synonymous with an upper hand in autonomy. As warfare is waged with machines – the smartest machines will have a formidable advantage. Machines that make immediate decisions and take immediate actions, even in situations where communications have been eliminated, will dominate machines with lesser intelligence. Machines that intelligently operate as highly coordinated fighting systems, sometimes called swarms, will surprise their enemies with emergent behaviors never witnessed. Any country that relies on Artificial Intelligence that relies on past examples to predict future events will be at an enormous disadvantage. The United States must continue to invest in Artificial Intelligence that integrates the domain expert (human) and the Machine Learning to move from extracting information from data to wisdom that enables actions. Brian Pierce, Deputy Director of DARPA's Information Innovation Office, has eloquently outlined a path forward in this regard. As machines interact with domain experts, creating a quid pro quo for each, the human experts' attention is directed to where the machine needs support, until we reach a point where the machine rarely needs support, and the machine will be able to make decisions and take actions when exposed to never observed situations and goals. Artificial Intelligence must be created in combination with human domain experts. In fact, the combination of artificial intelligence and quantum computing is another area that may bring game-changing capabilities to a country with an upper hand. The United States must stay at the forefront of creating exponential technologies that combine multiple emerging technologies.

In a digitally connected world and with the pervasiveness of social media and digital content, our foes can catalyze our self-destruction. Artificial Intelligence agents, armed with deep understanding of complexity theory, can create and distribute information that ignites internal battles and civil conflict. This is a subtle type of warfare that requires superior Artificial Intelligence to defend as compared the Artificial Intelligence needed to attack.

In the commercial world, we see the use of Artificial Intelligence accelerating design cycles and creating revolutionary designs in industries that relied for decades on evolutionary designs. As commercial assets are deployed, artificial intelligence is integrating information and actions from inspection, maintenance, and repair, and thus transforming traditional assets into "immortal machines." These immortal machines improve their performance over time, adapting to the current need or mission.

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Matthew Nielsen, Principal Scientist, Industrial Outcomes Optimization, GE Global Research

Questions submitted by Rep. Paul Tonko, House Committee on Science, Space, and Technology

Dr. Matt Nielsen, thank you for your work and for representing New York's Capital Region. I could not be more proud of the incredible research and work GE is doing at their Global Research facility in Niskayuna, New York. They have proven themselves time and time again as global leaders in the field.

Great technological advancements are also being made across New York's 20th District at RPI, SUNY Polytechnic Institute, the University of Albany, and at many other universities in our region. More and more, we are seeing data analytics and machine learning impact and improve countless industries and many aspects of our daily lives.

At SUNY Poly's state-of-the-art facilities, researchers from IBM, along with research alliance partners Global Foundries and Samsung, created a computer chip with transistors that are 5 nanometers wide, the smallest in the world. A computer chip the size of a fingernail can hold up to 30 billion of these transistors, which will mean faster, more powerful computing.

At UAlbany, The Albany Visualization and Informatics Lab specializes in data science and regional planning. One of their areas of focus involves the collection of large amounts of data on traffic patterns as a function of time during the day and year. Using these data, researchers develop models, test the validity of the model data, and predict traffic patterns and their correlation with weather patterns to direct commercial and public traffic more efficiently and better plan the development of new city or business districts.

Great strides are being made in this field within both the public and private sector. Continuing this trend will ensure safety and efficiency within all aspects of our lives.

Dr. Nielsen, it is important that we are encouraging future generations to engage in science and engineering fields to ensure that the important work you do is carried on.

1. **Does GE do any work in these areas with local universities and their students?**

Yes, the most recent example was announced a few weeks ago by RPI. Global Research is part of research team led by RPI that was awarded $1.4 million in project funding from Advanced Robotics for Manufacturing (ARM) to develop an advanced robotics solution in manufacturing that integrates sensors and automation technologies. The focus of the project is developing a mobile, or fixture-less robot-assisted platform that an operator could control and utilize in a manufacturing assembly process to improve productivity. Through the program, RPI students will have the opportunity to directly interact with GE researchers and other industrial partners on the project. In addition to funding from ARM, the program has received matching funds from the New York State Empire State Development Division on Science, Technology and Innovation (NYSTAR).

2. **What is GE doing to ensure that the next generation of scientist and engineers are prepared to enter the workforce to address these big data issues?**

One of the best ways we have engaged future scientists and engineers is through government R&D programs that bring industry and academic partners. We often will have the opportunity to work directly with engineering students on various projects. One example is through our involvement in the National Network of Manufacturing Innovation Institutes. The Innovation Institute for Additive Manufacturing, America Makes, has a great mix of industry, academic and other stakeholders that are focused on advancing additive, or 3D printing technologies. Managing data securely and reliably is a key consideration when dealing with digital files of a product or part design being sent to a 3D printer to be manufactured.

a. **Does GE engage with K-12 Capital Region students to encourage their interest in these fields?**

The key is reaching kids at a very early age during their primary and secondary education, which has been the focus of several local programs and events in the Capital Region that GE hosts throughout the year. Highlights include:

a. Science Day (4th grade students from three Capital Region schools – different schools selected every year) at Global Research in Niskayuna. Entering its 29th year this fall, GE invites 4th grade students from six area schools to see and experience firsthand the amazing way science impacts our world through more than a dozen experiments. Different schools are selected each year. In the nearly three decades GE has hosted the event, nearly 12,000 students from across the Capital Region have participated in Science Day;

b. GE Inspire Program with Schenectady High School – Inspire is a science and technology enrichment program at GE Global Research to expose intercity students to exciting cutting-edge technology that may inspire them to pursue a career in science or technology. The program is made up of three elements: technical concepts, non-technical concepts (communication skills), and a scientist mentor. During this 9- week program, students can learn about natural sciences, engineering and computer science. The Inspire program will enter its 10th year this fall;

c. GE Girls at RPI program – weeklong Summer STEM experience for local girls entering the eighth grade. The experience includes interactive and educational experiences in everything from wind turbine design and testing to robotics, chemistry, medical technologies and physiology and biomedical engineering. The program at RPI is one of several GE Girls programs that GE has supported across the nation;

d. Participating sponsor in the Niskayuna Central School District's annual Engineering Institute for Young Women, hosting interactive tours in robotics and other hands-on demonstrations at Global Research. This year was the 7th year we have been a sponsor of the Institute; and

e. First Robotics, Math Counts, Science Bowl and Invention Convention competitions… national competitions we sponsor and support locally in the Capital Region.

In addition to these formal events, Global Research scientists and engineers individually mentor students and participate in school events on their own time throughout the year.

Dr. Nielsen, in your testimony before the science committee on July 12th you discussed cyber-attacks and threats to power plants. This is important work. It is critical for our national security and our economy that such threats and attacks are stopped. You discussed GE's work on building the world's first "Industrial Immune System" for electric power plants that can detect and neutralize threats.

3. **Why is it so critical that we develop this type of "Industrial immune System"?**

In 2010 we witnessed a new era in cyber-attacks: Stuxnet. This attack was able to jump an air gap, get past informational and operational technology defenses, manipulate running control systems, and then execute a stealthy attack. This was the largest publicly documented attack and was focused on destroying industrial assets versus targeting the exfiltration of finance or personal information. Unfortunately, we are seeing an ever-increasing number of cyberattacks focused on industrial assets. What is clear, in the industrial space, is that we need to protect against adversarial nation states and well-organized hacker groups. This requires the continued research and development of more sophisticated defense technologies. The need is real and extremely critical, as the impact from a large scale cyber-attack focused on industrial assets could be devasting to both our citizens and economy.

4. **How effective is the "industrial immune system" in neutralizing threats? How fast does it work?**

While still in development, our goal is to detect and localize cyber-attacks with 99% accuracy. Once the industrial immune system has determined that a cyber-attack is present, we are targeting to provide neutralization for a majority of the system's critical functions, which have been compromised. Initial results provided by computer simulation studies and field data have indicated very good progress towards meeting these goals. The time requirements imposed on the industrial immune system are dictated by the dynamics of the system under protection. For our current focus on energy generation assets, the goal is to provide detection, localization and neutralization at the speed of the asset's control system, which is less than half a second.

5. **What are the next steps on this project? What work still needs to be done in this area by industry and the research community?**

To date, we have developed the key algorithms and validated them using sophisticated computer simulations and extensive field data. The key next steps and required investment include:

- Further research on neutralization for increased system resiliency during attacks; and
- Scaling and refining of the technology to provide protection for assets outside of electrical power generation sector.

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Matthew Nielsen, Principal Scientist, Industrial Outcomes Optimization, GE Global Research

Questions submitted by Rep. Jacky Rosen, House Committee on Science, Space, and Technology

As you all have discussed, scientists and companies continue to utilize big data analytics to better achieve research goals and improve industry needs. In Nevada, the Desert Research Institute – or DRI, the state's environmental research facility – uses extensive monitoring and modeling programs to analyze environmental and health data. For the past two years, DRI has been working with partners on the Healthy Nevada Project, one of the first community-based population health studies in the country. They are studying health, environmental, and socioeconomic data to better understand how these factors and genetics can help predict who may be at risk for certain diseases, allow for quicker diagnoses, and encourage the development of better treatments.

1. **In your view, is there a productive role that the Department of Energy can play in accelerating the development of technologies like those that the Healthy Nevada Project is using?**

Yes, DOE's high end computer processing capabilities will likely enable much faster analysis of the complex genomic, environmental and health data being collected in the Healthy Nevada Project. Their experience in big data modeling and analytics will also likely be complementary to standard statistical approaches. In principle, this means that answers, conclusions and recommendations should be arrived at more quickly over the course of the project, assuming sample size is sufficient and the correct data has been collected.

2. **How should we balance investments in DOE's computing facilities and advanced data analytics? Are we creating data faster than we can analyze it?**

The DOE delivers essential investment in advancing computational modeling and simulation, whereas investments from the commercial sector in scalable data analytics and machine learning remain comparatively robust. The growth of data is inevitably and inextricably entwined with the growth in the capabilities of digital technologies. It is quite true that the persistent acceleration in the world's capture and creation of data presents many challenges. These include the data must be stored cost-effectively, indexed to be usefully referenced and protected from unauthorized access or tampering. We can confidently argue it is also already true that without the assistance of computers such tasks would be impossible and people would be blind to the insights collected from these sensors, scientific instruments and computational models. Recognition of the commercial value in data has led to the success of some of today's most influential companies. To tame vast data, these companies have invested in developing scalable analytics, machine learning and data storage technologies now leveraged not only commercially, but also in science and engineering. Symbiotically, the advanced computing hardware in our Leadership Computing Facilities can also very effectively leverage scalable analytics, machine learning and storage technologies in conjunction with their time-proven capabilities in modeling and simulation. Modeling and simulation software and underlying software upon which it relies – such as mathematical libraries, optimized data structures and resource schedulers receives far more modest investment from commercial entities, however. Therefore, from the perspective of a user of both the computational methods and data analytics ecosystems, government emphasis on advancing the state of the art in modeling and simulation – and its ability to then exploit the commercial advances in scalable analytics and machine learning – would drive greater value in the combined ecosystem.

3. **How should the need to accelerate big data analytics and integrate private sector approaches influence the design requirements and success metrics of upcoming DOE computing acquisitions?**

DOE computing acquisitions should consider the complementary strengths in combining traditional physical modeling and simulation with scalable data analytics and machine learning methods. More powerful hardware and software will enable collection and creation of evermore complex and rich data. As we grow the scale, fidelity and multi-disciplinary nature of computational models, we will also greatly increase the synthetic data output from simulation of solid and fluid mechanics, thermodynamics, biochemistry, electromagnetics and other systems of study. A great variety of data already exceed the abilities of the human mind to comprehend. Computational tools will play an increasingly critical role to augment human perception and cognition through scalable analytics, machine perception and machine learning to improve the composition of highly-complex models and comprehension of their results. In turn, we can also employ modeling to improve both machine learning and data analytics by bounding results within the formalizations learned over the history of study of our scientific disciplines. In upcoming DOE computing acquisitions, and consistent with the CORAL procurement that resulted in the Summit machine at the Oak Ridge Leadership Computing Facility, recognizing these synergies between scalable analytics, machine learning and modeling and simulation in the requirements and success metrics will ensure the systems and the software will provide the most powerful methods as the state of the art of science and engineering advances through the collaboration of the human mind with the capabilities of computational methods and architectures.

Responses by Dr. Anthony Rollett

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Anthony Rollett, U.S. Steel Professor of Materials Science and Engineering, Carnegie Mellon University

Questions submitted by Rep. Gary Palmer, House Committee on Science, Space, and Technology

1. **In addition to serving here on the Energy Subcommittee, I also serve on the House Budget Committee where I am familiar with looking at government spending from a cost-benefit viewpoint. The President's budget request for FY 2019 included $899 million for the Advanced Scientific Computing Research Program, so I am wondering if you could speak a little more to the benefits that you see coming directly from that investment? In your opinion, what areas/technologies are giving us the best return on investment?**

Thank you for this question and for Congressional support for increased investment in this vital program. This increased funding supports a seamless web of capabilities all central to advancing the application of machine learning and artificial intelligence to manufacturing. It supports the supercomputing capabilities that support both material characterization research and the development and application of algorithms for real time analytics for materials and manufacturing process innovations. The investments, in both exascale computing and in developments in areas such as Quantum Computing, all contribute to U.S. leadership in this new frontier of manufacturing research and innovation.

A measure of return should clearly be evident in the straight line that may be drawn from these investments and the operation of associated user facilities to specific new breakthroughs in manufacturing related innovations in areas such as longer-lasting industrial components, new materials, more effective coatings, and improved data management and analytic capabilities.

I believe a critical need for the future will be to build even stronger linkages between high performance computing (HPC) and applications of machine learning and AI for manufacturing. As I noted in the hearing, we are at the early stages of this new paradigm in manufacturing research. Currently, our work is still largely utilizing algorithms developed for signal processing and information technology applications. Efforts to support focused development of machine learning and data analytics capabilities for advanced manufacturing will enhance the ROI in our manufacturing sectors from HPC programs.

In addition, opportunities for greater collaboration between materials scientists and computer scientists within HPC programs, including the potential for interdisciplinary center scale efforts, would increase the return on high performance computing investments for advanced manufacturing.

A second key strategy, and one that both the DOE labs and universities are embracing, is to aggressively develop and implement initiatives that engage small manufacturers and manufacturing related entrepreneurs in this research area, including targeted efforts to increase access to HPC facilities and interaction with lab and academic researchers.

2. **We all know that China is a major competitor in the machine-learning/AI space. What would it mean for the United States if another country were to gain dominance in machine learning?**

Thank you for the opportunity to comment on this vital issue. I approach this question not as an economist trained to assess specific economic impacts or outcomes but as a materials scientist and engineer who has spent much of his career in and around the workers, companies and communities that have been impacted by technological and international dynamics. I am confident that I speak for many of my research colleagues when I state that a key factor motivating our work has been the dream of advancing fundamentally new technologies that can change the dominant paradigm of the last several decades which has seen low cost labor in other nations reduce manufacturing opportunities for companies and workers in particular products and sectors.

The applications of machine learning and AI to manufacturing have the potential to accelerate entirely new approaches to manufacturing. They can particularly enhance applications such as additive manufacturing, which at scale could enable the cost effective production of radically new customizable, high value products that combine new materials as well as digital capabilities. Additive manufacturing is, by its very nature, a multiscale scientific challenge for developing the predictive capability that will allow engineers to take full advantage of these new technologies. It is already clear, for example, that co-design is essential, which means that part design must be done hand-in-hand with the design of the additive manufacturing process. The machine learning and AI applications are also central to the potential to integrate new production processes like additive manufacturing with robotics. These are the types of breakthroughs that we expect will facilitate new globally competitive manufacturing opportunities in the U.S.

Again, while I cannot project specific economic impacts or outcome, it would seem that as our workers, businesses and communities have endured the impacts of low cost competition from other nations in the past, it would be particularly unfortunate were we as a nation to lose leadership in these emerging high value manufacturing innovations.

There are also likely specific ramifications in the area of national security and defense should we lose leadership in these frontier innovations. Machine Learning and Artificial Intelligence are key enabling technologies for developing manufacturing breakthroughs that can shorten the defense supply chain and facilitate "in theater" production that can increase our warfighters' ability to flexibly respond to rapidly evolving threats from state and non-state actors. A balanced strategy for investment in the application of these technologies to manufacturing, which as I noted in my remarks we are only in the early stages of undertaking, will clearly support both sustained U.S. industrial and defense competitiveness.

3. **In your testimony you say that the importance of cybersecurity in manufacturing is not well understood. Can you give us some examples of potential negative consequences resulting from cyber-attacks on manufacturing operations? What kinds of steps are being taken to stay ahead of the curve on these attacks?**

Thank you for the opportunity to expand upon this point. In essence, the application of machine learning and artificial intelligence to advanced manufacturing will expand the "attack surface," the segment of our economy and production infrastructure vulnerable to cyber measures.

Scaling these applications involves intensive sensing at virtually all stages in the manufacturing process to generate the data that enable the ability to optimize operations and facilitate new product developments. Data collection, exchange and analytics activities will need to be cyber enabled within plants and across the supply chain to realize the full potential of the power of these applications. A few examples of vulnerabilities include: theft of intellectual property by interception of build files; disruption of builds by altering the build process; sabotage of parts by non-inspectable changes in part design; and reverse engineering of process design and part design by unauthorized monitoring of 3D printers.

This new manufacturing infrastructure is known as the industrial Internet of Things (IoT). Embedding cybersecurity objectives into the design and production of this IoT infrastructure will be vital. I know that the Science Committee has been a strong supporter of cybersecurity research—which now also includes the application of machine learning and artificial intelligence to cyber defense. In addition, the national manufacturing innovation institutes are examining opportunities to incorporate cybersecurity objectives in their missions to advance specific applications. The Department of Energy, in large part through research at national laboratories such as Idaho National Lab, has been focused on the security of sensor and computer control devices. While initially targeted to enhance protection of the grid, these investments can enhance the application of cybersecurity technologies to advanced manufacturing.

Speaking as a materials scientist and engineer, I also would strongly encourage programs and research environments that foster collaboration across advanced manufacturing and cybersecurity disciplines. One effective strategy at Carnegie Mellon's cybersecurity research and education institute, CyLab, has been to create an environment that fosters dynamic interaction between "makers" and "breakers," i.e., direct collaboration between those researchers advancing new digital applications and products and those with knowledge of hacking and cyber-attack capabilities. These kinds of dynamic collaboration models will be important to bring a focus on cybersecurity into the heart of advanced manufacturing research.

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

"Big Data Challenges and Advanced Computing Solutions"

Dr. Anthony Rollett, U.S. Steel Professor of Materials Science and Engineering, Carnegie Mellon University

Questions submitted by Rep. Jacky Rosen, House Committee on Science, Space, and Technology

As you all have discussed, scientists and companies continue to utilize big data analytics to better achieve research goals and improve industry needs. In Nevada, the Desert Research Institute – or DRI, the state's environmental research facility – uses extensive monitoring and modeling programs to analyze environmental and health data. For the past two years, DRI has been working with partners on the Healthy Nevada Project, one of the first community-based population health studies in the country. They are studying health, environmental, and socioeconomic data to better understand how these factors and genetics can help predict who may be at risk for certain diseases, allow for quicker diagnoses, and encourage the development of better treatments.

1. **In your view, is there a productive role that the Department of Energy can play in accelerating the development of technologies like those that the Healthy Nevada Project is using?**

While health and environmental research is not my field, I commend the creative and innovative application of data analytics and machine learning that is being undertaken by the Desert Research Institute. This capacity for machine learning to have a transformative impact in the very understanding of fundamental issues is what I and my colleagues in materials sciences experience in manufacturing.

It also strikes me as the kind of innovative compelling application that can attract and encourage students to consider education and careers in computational fields and areas of study---a vital need for our nation.

Additionally, I would expect the DRI, as an educational institution, to be interested in participating in the many opportunities for scientific research offered by the DOE.

This effort seems very similar to an initiative I noted in my testimony, the partnership between DOE and the Veterans Administration to apply AI and DOE's high performance computing capabilities to the health challenges of Veterans. As with DRI's efforts, this initiative is targeting opportunities to apply DOE's resources and growing AI capabilities to high impact health related challenges.

2. **How should we balance investments in DOE's computing facilities and advanced data analytics? Are we creating data faster than we can analyze it?**

Thank you. This question in part speaks directly to a key challenge for the future of machine learning and artificial intelligence applications for manufacturing: the need to balance and integrate the utilization of high performance computing with cloud computing. Both will be vital to advancing key applications to accelerate new material development and improvements in manufacturing processes made possible by digital engineering. DOE should lead the field in blending these fields, in collaboration with the NSF, DoD and NASA. Supporting and encouraging the creation of interdisciplinary teams that bring together domain experts in the fields of materials science with computer scientists, data analytics researchers, as well as social scientists in the fields of privacy and ethics, will be vital to develop new applications of machine learning and AI tools for manufacturing and should become a design feature of future DOE programs.

This interdisciplinary approach will also be essential to address the challenges posed by the explosion of data referenced in your question. We are only at the beginning of the revolution in connecting physical systems in manufacturing and infrastructure to the digital world. As I noted in my testimony, machine learning is already proving vital for accelerating the conversion of data into useful information. Advances in raw computing speed and computable problem size are still crucial to many national needs in defense, transportation safety, climate change and many other areas.

In addition to advances in computer science, data analytics, and stronger computational foundations in fields such as materials science, attention to the vital dynamics of human/computer interaction and teaming will be essential across the educational spectrum as jobs and a wide variety of daily interactions will increasingly seek to leverage digital intelligence to augment human creativity.

3. **How should the need to accelerate big data analytics and integrate private sector approaches influence the design requirements and success metrics of upcoming DOE computing acquisitions?**

While I have benefited enormously from the ability to utilize DOE user facilities, including advanced computing resources, I am not qualified to contribute specific recommendations on metrics relating to the design and acquisition of DOE computing capabilities. The inherent reference in this important question to the accelerating pace and scale of breakthroughs in private industry captures a dynamic that is shaping innovation across a number of fields and machine learning and AI in particular.

I should take the opportunity to commend the manifold channels through which the Federal government seeks input from the scientific community, notably the Federal Advisory Committee system, of which the DOE makes excellent use. Research and educational initiatives that strongly encourage and directly foster closer collaboration among lab scientists, universities, industrial companies and emerging technology companies, which several DOE labs are pursuing, will be increasingly valuable to address and capture the powerful benefits of this trend for agency and national research missions.

Appendix II

Additional Material for the Record

DOCUMENTS SUBMITTED BY REPRESENTATIVE NEAL DUNN

CNN Money Live TV

MIT scientists created an AI-powered 'psychopath' named Norman

By Sara Ashley O'Brien
Updated 4:21 PM EDT, Thu June 07, 2018

(CNNMoney) - Norman always sees the worst in things.

That's because Norman is a "psychopath" powered by artificial intelligence and developed by the MIT Media Lab.

Norman is an algorithm meant to show how the data behind AI matters deeply.

MIT researchers say they trained Norman using the written captions describing graphic images and video about death posted on the "darkest corners of Reddit," a popular message board platform.

The team then examined Norman's responses to inkblots used in a Rorschach psychological test. Norman's responses were compared to the reaction of another algorithm that had standard training. That algorithm saw flowers and wedding cakes in the inkblots. Norman saw images of a man being fatally shot and a man killed by a speeding driver.

"Norman only observed horrifying image captions, so it sees death in whatever image it looks at," the MIT researchers behind Norman told CNNMoney.

Named after the main character in Alfred Hitchcock's "Psycho," Norman "represents a case study on the dangers of Artificial Intelligence gone wrong when biased data is used in machine learning algorithms," according to MIT.

We've seen examples before of how AI is only as good as the data that it learns from. In 2016, Microsoft launched Tay, a Twitter chat bot. At the time, a Microsoft spokeswoman said Tay was a social, cultural and technical experiment. But Twitter users provoked the bot to say racist and inappropriate things, and it worked. As people chatted with Tay, the bot picked up language from users. Microsoft ultimately pulled the bot offline.

The MIT team thinks it will be possible for Norman to retrain its way of thinking via learning from human feedback. Humans can take the same inkblot test to add their responses to the pool of data.

According to the researchers, they've received more than 170,000 responses to its test, most of which poured in over the past week, following a BBC report on the project.

MIT has explored other projects that incorporate the dark side of data and machine learning. In 2016, some of the same Norman researchers launched "Nightmare Machine," which used deep learning to transform faces from pictures or places to look like they're out of a horror film. The goal was to see if machines could learn to scare people.

MIT has also explored data as an empathy tool. In 2017, researchers created an AI tool called Deep Empathy to help people better relate to disaster victims. It used technology to visually simulate what it would look like if that same disaster hit in your hometown.